Acc. No. 15358

Contents

Preface, xxix

PART ONE: INTRODUCTION

1

The Economic Approach, 1

Economic Systems, 1
Scarcity and Choice, 2 Economic Goals, 3 Positive and Normative Economics, 4

Economic Analysis, 7
Early Economic Thought, 7 Classical Economics, 8 Neoclassical Economics, 8 The Literature, 12 Schools and Groups, 13 The Economic Approach: Matters of Logic and Degree, 13

Summary, 14
Key Concepts, 15
Questions for Review, 15

2

Basic Economic Principles, 17

The Economy as a System, 18

Households: The Decision Units for Selling Inputs and Buying Goods, 19 Enterprises and Inputs, 20 Maximizing Behavior, 21 Market Exchange, 22 The Circular Flow, 23

Microeconomic Principles, 24

Opportunity Cost, 25 Marginal Conditions, 25 Diminishing Marginal Effect, 26 Efficiency and Scarcity: The Production Possibility Boundary, 27 Equilibrium, 31 Public Choice, 32

Macroeconomic Principles, 33

The Circular Flow, 33 Income and Output, 34 Demand and Income, 34 Changes in Output, 35 Potential Output, 35 Stabilization Policy, 36

Summary, 37
Key Concepts, 37
Questions for Review, 37

338.5
Ar23M

3

Methods and Measurements, 39

Diagrams and Their Use, 40

Linear Equations and Their Graph, 41 Economic Models, 44 Time Series, 48 Distributions, 49

The Interpretation of Numbers, 50

Problems of Bias and Deception, 50 Stocks and Flows, 51

Summary, 53
Key Concepts, 54
Questions for Review, 54

PART TWO: MICROECONOMICS

4

Demand and Supply, 55

Demand, 58

Influences on Demand, 58 The Demand Curve, 59 Quality Demanded and Demand, 61 Price Elasticity of Demand, 63 Elasticity and Total Revenue, 66 Determinants of the Price Elasticity of Demand, 67 Income Elasticity of Demand, 67 Cross-Elasticity of Demand, 70

Supply, 71

Influences on Supply, 71 The Supply Curve and Its Upward Slope, 72 Quantity Supplied Versus Supply, 73 Elasticity of Supply, 75 Determinants of Elasticity of Supply, 77

Interaction of Supply and Demand, 78

Summary, 81

Key Concepts, 83

Questions for Review, 83

330
S31E

5

Demand and Supply in Action, 85

Effects of Elasticities on Market Outcomes, 86

Elastic Demand and Supply, 86 Inelastic Demand and/or Supply, 87 The Price of Oil, 90

Interferences with the Market Process, 91

The Burden of a Sales Tax, 91 Price Controls, 94 Controls on Quantity, 97

Measuring Supply and Demand, 98

Summary, 101

Key Concepts, 102

Questions for Review, 102

6

Individual Demand, 103

The Analysis of Utility and Demand, 104

Rational Choices by Consumers, 105 Diminishing Marginal Utility, 107 Individual Demand Curves: Preferences and Income, 110 Scarce Goods and Free Goods, 111 Marginal Utilities and Prices in Equilibrium, 113 Shifts Versus Movements Along Demand Curves, 116 Consumer Surplus, 117 Market Demand Is the Sum of Individual Demands, 118 Derived Demand, 120

The Validity of Demand Analysis, 120

Summary, 121

Key Concepts, 122

Questions for Review, 122

7

The Enterprise, 123

Patterns of Actual Enterprises, 124

Private Enterprises, 124 Other Types of Firms: Public Enterprises, Nonprofit Firms, and Cooperatives, 135

The Enterprise, 135

Choices and Outcomes, 136 Inputs, Outputs, and Production, 137 Simple Accounting, 138 Success Indicators: Profitability and Stock Prices, 141

A Case Study: Starting a New Enterprise, 142

Choosing What to Produce, 142 Starting, 143 Lessons, 144

Summary, 145

Key Concepts, 145

Questions for Review, 146

◂ 8 ▸

Supply: The Nature of Costs, 147

Basic Concepts: Technology, Opportunity Cost, and Economic Profit, 148

Technology, 148 Opportunity Cost, 149 Economic Profit, 151

Productivity and Costs in the Short Run, 153

Short Run and Long Run, 153 Productivity in the Short Run, 153 The Law of Diminishing Marginal Returns, 157 Costs in the Short Run, 158

Productivity and Costs in the Long Run, 163

Derivation of the Long-Run Average Total Cost Curve, 164 Economies of Scale: The Shape of the Long-Run ATC Curve, 167 The Marginal Conditions Necessary for Least-Cost Production, 171

Summary, 173

Key Concepts, 174

Questions for Review, 175

◂ 9 ▸

Pricing and Output Under Perfect Competition, 177

The Rules for Maximizing Profits, 178

Should the Firm Produce at All? 178 How Much Should the Firm Produce? 178

Setting Output Under Perfect Competition, 180

The Nature of Pure Competition, 180 Competition: A Process and a Zone of Choice, 180 Rivalry and Pure Competition May Give Similar Results, 181 Firm and Market Demand in Perfect Competition, 181 Marginal Cost, 182 The Firm's Short-Run Supply Curve in Perfect Competition, 183 A More Complete Analysis, 184 The Short-Run Market Supply Curve Is a Summation, 188 Shifts in the Firm and Industry Supply Curves, 189 Short-Run and Long-Run Equilibrium in Perfect Competition, 190

Long-Run Supply, 193
The Firm's Long-Run Supply Curve, 193 The Long-Run Market Supply Curve, 194

Efficient Allocation Under Competition, 194
Summary, 196
Key Concepts, 197
Questions for Review, 197

10

Monopoly, 199

Varieties of Monopoly and Competition, 200
Monopoly and Its Effects, 200
The Characteristics of Monopoly, 200 How Monopoly Power Is Created and Maintained, 203 The Effects of Monopoly Power, 204 Price Discrimination, 213

Cases of Monopoly, 216
Standard Oil, 217 Electric Companies, 217

Summary, 218
Key Concepts, 218
Questions for Review, 219

11

Degrees of Competition, 221

The Dominant Firm, 222
Definition of Dominance, 222 Instances and Effects of Dominance, 223 Possible Causes of Dominance, 224

Oligopoly, 229
Concentration and Leading Firms, 229 Conflicting Incentives Can Make Oligopoly Unstable, 232 Types of Collusion, 234 The Central Tendency Under Oligopoly, 235 Rigid Prices: Kinked Demand Curves, 236 Economies of Scale: A Cause of Oligopoly? 239

Monopolistic Competition, 240
Patterns and Trends in Real Markets, 241
Aggregate Concentration, 241 Concentration in Individual Markets, 242 Conglomerate Firms, 244

Summary, 246
Key Concepts, 247
Questions for Review, 247

◂ 12 ▸

Policies Toward Monopoly Power: Antitrust, 249

Origins and Standards of U.S. Antitrust Policies, 250
Three Waves, 250 Standards of Efficient Policies, 252

U.S. Antitrust: Forms and Coverage, 253
The Agencies and Laws, 253 History, 254 Antitrust Criteria, 256 Precedents, 257 Mergers, 258 Economic Effects, 258

Specific Parts of Antitrust, 259
Antitrust Actions Toward Existing Concentration, 259 Antitrust Policies Toward Mergers, 264 Policies Toward Price Fixing and Other Actions, 267

Summary, 269
Key Concepts, 270
Questions for Review, 270

◂ 13 ▸

Policies Toward Monopoly Power: Regulation and Public Enterprise, 271

Regulation of Utilities, 272
Patterns of Regulation, 272 Decisions on Price Levels and Structures, 275 Four Economic Issues of Regulation, 277

Public Enterprise, 283
Coverage and Purposes, 283 Subsidies and Efficiency, 286

Summary, 287
Key Concepts, 287
Questions for Review, 287

14

Input Markets, 289

The Demand for Inputs, 290
The Level of Input Use, 290 Marginal Revenue Product, 290 The Profit-Maximizing Level of Input Use, 292 Deriving the Firm's Demand Schedule for an Input, 293 Elasticity of Demand for an Input, 294 Shifts in the Marginal Revenue Product Schedule, 295 Comparison of Input Use of a Perfect Competitor and a Firm with Monopoly Power, 295

Supply and Equilibrium in Input Markets, 297
The Supply of Inputs, 297 Economic Rent, 299 Taxing Economic Rents, 301 Who Provides the Value of Production? 302

Summary, 303
Key Concepts, 304
Questions for Review, 304

15

The Economics of Labor and Unions, 305

Labor Supply, Demand, and Market Outcomes, 306
The Marginal Utility of Work, 305 The Choice of a Job, 307 Individual Labor Supply Schedules, 308 Price and Income Effects, 308

The Market Supply Curve of Labor, 309 The Demand for Labor, 311 Equilibrium Between Supply and Demand, 313

Differences in Labor Skills, 315
Variations in Pay Rates, 315 Investment in Human Capital: The Cost of Training, 316 Scarcity of Talent, 319

Departures from Competitive Market Outcomes, 320
Monopsony and Competitive Supply, 325 Bilateral Monopoly, 325 The Effects of Labor Groups on Workers' Incomes, 326

Summary, 328
Key Concepts, 329
Questions for Review, 329

16

Capital, Investment, and Technological Change, 331

Capital and Investment Decisions, 332
What Is Capital? 332 Actual Capital and Investment, 332 The Decision to Invest, 333 Market Demand and Supply for Capital, 337

The Return to Capital, 338
Risk and Return, 338 Interest, 341 Profit, 342

The Value of Assets, 343
Fluctuations in Asset Values, 343 Expectations Govern Asset Values, 346 Bonds and Stocks, 347 The Choice Process Equalizes Returns, 348 Three Levels of Knowledge, 348 Stock Markets as a Control System, 348

Capital and Technological Change, 349
Trends of Capital and Productivity, 349 Forms and Components of Technological Change, 351 Decisions to Innovate, 352 The Patent System, 353

Summary, 353
Key Concepts, 354
Questions for Review, 354

◂ 17 ▸

General Equilibrium, 355

The General Process Toward Equilibrium, 356

Conditions and Processes of General Equilibrium, 358

The Conditions, 358 Ripple Effects, 361

Limits on Competitive Efficiency, 364

External Costs and Benefits, 364 Distribution May Be Unfair, 367
Cultural Values Are Not Necessarily Provided, 368
Monopoly, 368 Natural Resources, 369

Summary, 370
Key Concepts, 370
Questions for Review, 370

◂ 18 ▸

Public Finance, 373

Economic Concepts of Optimal Public Policies, 374

Social Goods, 374 External Effects, 377 Public Expenditure: Cost-Benefit Analysis, 378 Cost-Benefit Analysis for Specific Projects, 380 Alternatives to Public Spending and Taxes, 383 Categories of Spending and Taxes, 385

Taxes: Impacts on Distribution and Incentives, 386

Incidence: Analyzing Who Bears the Burden of Taxes, 386 Incentives: How Taxes May Affect Choices, 387 Distribution: The Effects of Taxes and Spending, 389

Major Patterns of Public Finance, 392

Trend and Share, 392 Composition: Local, State, and Federal Shares, 393 Purchases and Transfer Payments, 394 The Variety of Spending Programs, 395 Taxes, 396

Summary, 398
Key Concepts, 398
Questions for Review, 399

19

Inequality, Poverty, and Discrimination, 401

Income Differences and Their Causes, 402

The Degree of Inequality, 402 Technical Causes of Apparent Inequality, 404 The Economic Forces Shaping the Income Distribution, 405

Discrimination, 406

Employment Discrimination, 406 Discrimination in Housing, 408

Public Policy and Income Distribution, 409

Tax Policies, 410 Public Expenditures, 413 Equal Opportunity Programs, 413 Minimum Wage Laws, 416

Summary, 417
Key Concepts, 418
Questions for Review, 418

20

Education, Social Regulation, and the Military, 419

The Economics of Education, 420

Private Benefits of Education, 420 Public Benefits of Education, 421 Actual Expenditures on Education, 422 Public Schools and the Issue of Choice, 423 Financing Public Colleges: Efficient? Fair? 425

Social Regulation: Protecting the Environment, Workers, and Consumers, 428

Environmental Issues and Programs, 428 Cost-Benefit Issues, 430 The Use of Rules to Limit Pollution, 432 The Use of Incentives, 433 Programs Protecting Worker and Consumer Safety, 435

Military Spending, 437

Avoiding Waste in Producing Military Goods, 438 Efficient Military Levels and the Arms Race, 440 The Economic Basis for a Volunteer Army, 442

Summary, 445
Key Concepts, 446
Questions for Review, 446

◂ 21 ▸

Natural Resources: Concepts and Policies, 447

Basic Concepts, 448

Conservation: Reaching the Optimum Rate of Use, 448 Five Determinants of the Optimum Rate, 452 Private Markets Can Optimize the Use of Resources, Except . . ., 455

Agricultural Economics, 457

Basic Conditions, 457 Farm Policies, 459

The Economics of Energy, 462

Basic Trends, 462

Future World Resources, 465
Summary, 466
Key Concepts, 467
Questions for Review, 467

PART THREE: MACROECONOMICS

22

An Introduction to Macroeconomic Analysis, 469

Long-Term Trends, 1870–1980, 470

Gross National Product, 470 Unemployment, 471 Consumer Prices, 472 Common Stock Prices, 473 The Federal Government, 474

Cyclical Fluctuations in the 1970s, 477

Production, 477 Unemployment, 478 Consumer Prices, 479 Common Stock Prices, 479 Macroeconomics and the 1970s, 480

Summary, 481
Key Concepts, 481
Questions for Review, 481

23

National Product and Income, 483

The Circular Flow of Goods and Services, 485

A Two-Sector Economy, 485 A Four-Sector Economy, 486

Measuring National Output and Income, 488

The Gross National Product, 488 Input-Output Accounting, 489 Value Added and Gross National Income, 491 Deliveries to Final Demand, 492

Sectoral Surpluses and Deficits, 495

Net Taxes, Business Saving, and the Distribution of GNY, 495 Sectoral Receipts and Expenditures, 497

Measuring Trends in Prices and Output, 498
Price Indexes, 499 Real GNP and the GNP Deflator, 501

Summary, 502
Key Concepts, 503
Questions for Review, 503

◂ 24 ▸

Equilibrium of the Circular Flow, 505

Consumption and Equilibrium, 506
The Propensity to Consume, 506 Combining the Two Sectors, 510 Saving, Investment, and the Sectoral Surpluses, 511

Income and Spending of the Business Sector, 513
Variations in Business Saving, 513 Variations in Planned Investment, 514 Equilibrium and the Business Deficit, 515

Government, Foreign Trade, and Equilibrium, 516
The Government, 516 Foreign Trade, 517
Equilibrium of All Four Sectors, 518

Summary, 520
Key Concepts, 520
Questions for Review, 520

◂ 25 ▸

The Multiplier, 523

The Multiplier Theorem, 525
The Multiplier in Action, 525 Why Does the Multiplier Work? 526

The Multiplier Process, 528
Actions and Reactions, 529 The MDP, 530 A More Realistic Example, 531 The MDP and the Size of the Multiplier, 531 The Multiplier and Stability, 533 Factors Affecting the Size of the Multiplier, 534

The Uses, Misuses, and Limits of the Multiplier, 535
Summary, 537
Key Concepts, 537
Questions for Review, 538

26

Unemployment and Inflation, 539

The Cost of Inputs, 540
The Cost of Intermediate Goods, 540 Imports, 542 Direct and Indirect Labor Costs, 544 The Importance of Labor Unions, 544 The Phillips Curve, 546 Why Is There a Phillips Curve? 548

Productivity and Cost of Output, 549
The Arithmetic of Costs, 549 Productivity, Costs, and Prices, 550 The Supply Side: A Wrap-Up, 551

Inflation and the Demand for Goods, 552
Categories of Unemployment, 552 GNP, Unemployment, Wages, and Prices, 554 Prices and Demand, 556

Persistent Inflation, 557
What Happened in the 1970s? 557 Price, Wages, and Expectations, 558 The Shifting Phillips Curve, 560

Summary, 562
Key Concepts, 563
Questions for Review, 564

◂ 27 ▸

Financing the Circular Flow, 565

Money, 566

The Functions of Money, 567 The U.S. Money Supply, 568 Money as Debt, 569

Financial Institutions and Assets, 570

Financial Assets, 570 Commercial Banks, 571 Savings Institutions, 572 The Security Markets, 572

The Flow of Funds, 573

Financial Sectors and Production Sectors, 573 The U.S. Flow of Funds Account, 573

Summary, 574

Key Concepts, 575

Questions for Review, 575

◂ 28 ▸

Banks and Money Creation, 577

The Institutions of the Banking System, 578

Commercial Banks, 578 Check Clearing, Bank Reserves, and Reserve Requirements, 579 The Depository Institutions Deregulation and Control Act, 581 Commercial Bank Assets and Liabilities, 582 Federal Reserve Assets and Liabilities, 583

The Creation of Bank Money, 584

Bank Lending and Deposit Creation, 584 Multiple Deposit Creation, 585 Deposit Contraction, 588 Some Complications, 589

The Federal Reserve and Money Creation, 591

The Monetary Base, 591 Open Market Operations, 592 Changes in Reserve Requirements and the Fed's Lending Rate, 593 The Fed's Operations as a Whole, 594

Summary, 594
Key Concepts, 595
Questions for Review, 595

◂ 29 ▸

Money, Interest, and GNP, 597

The Determination of Interest Rates, 598
The Rate of Interest, 598 Money Demand, Interest, and Velocity, 599 The Equilibrium Rate of Interest, 604 Changes in the Interest Rate, 606 The Federal Reserve and the Rate of Interest, 606

Interest and Expenditures, 608
Consumer Demand, 608 Residential Investment and the Interest Rate, 608 Business Investment, 609

Combining the Markets, 611
The Monetary Feedback, 611 The Federal Reserve and GNP, 612 Keynesianism and Monetarism, 614 Money and Inflation, 617 Money and Interest in the Long Run, 619

Summary, 620
Key Concepts, 621
Questions for Review, 621

◂ 30 ▸

The Institutions, Goals, and Strategies of Stabilization Policy, 623

The Conduct of Fiscal Policy, 624
The Employment Act of 1946 and the Council of Economic Advisers, 624 The Office of Management and Budget and the Cabinet-Level Departments, 625 The Congress, 625

The Goals of Stabilization Policy and the Costs of Instability, 625
The 1962 Economic Report, 626 Policy Guidelines, 626 Potential GNP and the Gap, 627 The Gap as Lost Output, 629 The Gap as Wasted Resources, 629 The Cost of Exceeding Potential Output, 630 Stopping Persistent Inflation: The Costs, 630 Living with Persistent Inflation: The Costs, 632 Price and Wage Controls, 635

Strategies of Stabilization, 636
Monetarism and Monetary Rules, 636 Discretionary Stabilization Policy, 638

Summary, 639
Key Concepts, 640
Questions for Review, 640

◂ 31 ▸

Fiscal and Monetary Policy, 641

The Principles of Fiscal Policy, 642
The Federal Budget, 642 The Multiplier Effects of Changes in the Federal Budget, 643 The Balanced Budget Multiplier, 644 Limitations on the Multiplier Effects of Budget Changes, 644 The Built-In Stabilizers, 645 The High-Employment Budget, 646

The National Debt, 649
Who Owns the Debt? 650 The Burden of Debt, 651

The Conduct of Monetary Policy, 652
Open Market Policy, Reserve Requirements, and Discounting, 653 Control Over the Money Supply, 654 Control Over Interest Rates, 655

Policy Coordination and Conflict, 656
National Defense Spending, 657 Financing Transfer Payments, 659

Summary, 660
Key Concepts, 661
Questions for Review, 661

32

Stabilization Policy: The Historical Record, 663

The Early 1960s, 664

Symptoms: Weak Recoveries from the Recessions of 1958 and 1961, 664 Diagnosis: Business Investment and the Federal Budget, 665 The Prescription: A Tax Cut, 667

The Late 1960s, 669

The Credit Crunch of 1966, 669 The Tax Surcharge of 1968, 671

The Nixon-Ford Years, 1969–1976, 672

The Business Cycle in the Late 1960s and the 1970s, 672 Fiscal and Monetary Policies During the First Administration, 673 The Period of Direct Controls, 674 The Second Republican Administration, 676

The Carter Administration, 676

Prices and Employment Under the Carter Administration, 677 Fiscal Policy, 678 Monetary Policy, 679 Was Stabilization Policy a Failure During the Carter Years? 680

The Reagan Program, 681

Supply-Side Economics, 681 The Defense Budget, 683

Postscript on Policy Making, 683

Summary, 683

Key Concepts, 684

Questions for Review, 684

33

American Economic Growth, 685

Population, 688

Population and Immigration, 688 Internal Migration, 692 Urbanization, 693 Other Demographic Changes, 695

Capital Accumulation and Technical Change, 699
The Theory of Growth, 699 Productivity in Agriculture, 700 The Growth of Manufacturing, 702 Transportation, 706

Summary, 711
Key Concepts, 712
Questions for Review, 712

PART FOUR: INTERNATIONAL ECONOMICS

◂ 34 ▸

International Trade, 713

Comparative Advantage and the Benefits of Trade, 714
Comparative Advantage, 714 Specialization and Gains from Trade, 715 Beyond Constant Cost, 717 Many Goods and Many Countries, 718 The Pattern of World Trade, 719

Protectionism, 720
Tariffs and Quotas, 721 The Case for Protection, 724 Export Restrictions, 726

Foreign Trade and the U.S. Economy, 727
The Growing Importance of Trade, 727 The Composition of Trade, 731

Summary, 734
Key Concepts, 734
Questions for Review, 735

◂ 35 ▸

International Finance, 737

International Currencies and Payments, 738

Currency Markets, 739 The Balance of Payments, 740 The Current Account, 743 Changes in GNP, 745 Changes in Relative Prices, 745 The Capital Account, 747 Direct Investment, 751 Portfolio Investment, 752

The World Payments System, 753

The Gold Standard, 753 The Bretton Woods System, 755 Payments Equilibrium Under the Bretton Woods System, 758 The Demise of the Bretton Woods System, 759

Summary, 762
Key Concepts, 763
Questions for Review, 764

◂ 36 ▸

Economic Development, 765

What Are the Third World Economies Like? 767

The Dimensions of Poverty, 767 Industrial Structure, 772 Uneven Development, 774

Why Are They Underdeveloped? 774

Capital, Saving, and Growth, 775 Population Growth, 777 Education and Development, 778 Agriculture and Development, 779 Relationships with the First World, 780

How Can They Develop? 781

Capitalism and Socialism, 781 Capital Accumulation, 782 Population Control, 783 Agriculture, 784 Trade and Development, 785 Foreign Capital, 786

Summary, 788
Key Concepts, 789
Questions for Review, 789

PART FIVE: MARXISM AND SOCIALIST ECONOMICS

37

Marxism and Capitalism, 791

Work, Value, and Surplus Value, 792

Commodities, 793 Direct and Indirect Labor, 793 Labor, Labor Power, and Surplus Value, 794 Determinants of Surplus Value, 796

The Accumulation of Capital, 797

The Class Structure of Capitalism, 797
The Reserve Army of the Unemployed, 799
The Reserve Army, the Accumulation of Capital, and Periodic Crises, 800 The Concentration of Capital and the Growth of Unproductive Labor, 800
Foreign Trade and Investment, 802
Contradictions and the Collapse of Capitalism, 803

Summary, 804
Key Concepts, 804
Questions for Review, 805

38

The Economies of Russia and China, 807

Economic Planning: The Soviet Economy, 808

Some Historical Background, 808 Stalin's Development Strategy, 809 Soviet Economic Structure, 811 The Planning Process, 813 Prices, 815 Income Distribution, 816 Is Russia a Socialist Country? 817

The People's Republic of China, 817
The Chinese Revolution, 818 Reconstruction (1949–1952) and the First Five-Year Plan (1952–1957), 819 The Great Leap Forward 1958–1960, 821 China Since the Great Leap, 823 The Future of China, 824

Summary, 825
Key Concepts, 825
Questions for Review, 825

GLOSSARY, 827
INDEX, 843

diminishing marginal effect, scarcity, equilibrium, public choice). Macroeconomic principles.

Chapter 3 The linkage of diagrams, hypotheses, and models. Organizing and presenting data. Stocks and flows.

Part II presents allocation analysis thoroughly and with unusual concern for general equilibrium.

Chapter 4 A focus on supply and demand concepts. Thorough treatment of elasticity.

Chapter 5 Supply and demand concepts and cases are knitted together; agriculture, oil prices, tax incidence, market controls. Measures of elasticities.

Chapter 6 Focuses on individual demand. Consumer surplus. An assessment of utility analysis.

Chapter 7 Uniquely thorough coverage of the enterprise. Actual patterns of firms. A tour of *Wall Street Journal* data. Nonprivate firms. Accounting, motives, and success indicators. A case study of starting up a firm.

Chapter 8 Intensive coverage of cost analysis. Link between productivity and cost. Economies of scale.

Chapter 9 The nature of competition, marginal cost, and efficiency conditions.

Chapters 10–13 cover monopoly power and its policy remedies with an industrial-organization focus.

Chapter 10 Varieties of markets, from pure monopoly to pure competition. Concise causes and effects of monopoly. Case studies of monopolies. Price discrimination fully explained.

Chapter 11 The dominant-firm case. Numerous practical instances, including newspaper markets. The Schumpeterean process. The contrast between tight and loose oligopoly. New data on the rise in competition since 1960.

Chapter 12 Antitrust agencies, trends, and criteria. A detailed presentation of cases and their economic effects.

Chapter 13 The economic content of regulation. Commissions and their setting. Key economic issues: marginal cost pricing, inefficiency, and deregulation. Public enterprise: its coverage and economic criteria.

Next come inputs and general equilibrium in Chapters 14–17. Both labor and capital are given detailed attention.

Chapter 14 Thorough analysis of input choices. Economic rent. Inputs' roles in creating value.

Chapter 15 The utility basis of work choices. Human capital and returns to education. Effects of labor unions.

Chapter 16 Uniquely thorough, integrated coverage of capital (physical and portfolio) and technological change. Investment choices, cost of capital, return to capital; risk and asset values. Expectations and stock prices. Stock markets as the control system of capitalism. Trends and elements of technological change.

Chapter 17 Complete coverage of equilibrium and allocation. Ripple effects and input-output tables. Limits on the invisible hand.

We round out microeconomics by analyzing major public policy choices in Chapters 18–21.

Chapter 18 A thorough analysis of social goods, external effects, and cost-

benefit analysis. Taxes and incentive effects. Trends of taxes and spending.

Chapter 19 Trends and causes of inequality. Analysis of discrimination. Actual incidence of taxes and spending.

Chapter 20 Analysis of resource conservation: criteria and free-market efficiency. Common-property resources. Agriculture, the energy sector, and future world resource scarcities.

Chapters 22–33 survey macroeconomic and monetary analysis. They contain a thorough, careful exposition of the standard theoretical tools, extensive historical instances of their application, and numerous unique features.

Chapter 22 Trends in output and prices. Fluctuations in employment. The growing federal government share. Stock prices and the business cycle. The pervasiveness of cyclical fluctuations.

Chapter 23 Visualizing the economic system as a whole. National accounts. Value added, final demand, and gross national income. How the input-output structure ties industries together. Sectoral receipts and expenditures. Price indexes and real output.

Chapter 24 The consumption function. Planned and unplanned investment. Business saving. Equilibrium in a two-sector economy. How the government and foreign sectors fit in.

Chapter 25 The multiplier as theorem, process, and number. Expansion and contraction in a two-sector economy. The four-sector economy. Limits, uses, and misuses of the multiplier.

Chapter 26 Cost push, input prices, and productivity. Final demand, GNP, and input markets. The Phillips curve. Persistent inflation. Expectations. The shifting Phillips curve. (The role of money in the inflationary process is deferred to Chapter 29.)

Chapter 27 An introduction to money and monetary institutions. Definitions of money. Financial assets. Commercial banks, thrift institutions, and security markets. The flow of funds.

Chapter 28 The creation of bank money. The Federal Reserve and the banking system under the 1980 banking law. Assets and liabilities of the Fed and the banking system. How banks expand and contract the money supply. Open market operations and their impact on the monetary base. Reserve requirements. Federal Reserve lending.

Chapter 29 The level and structure of interest rates. How interest rates affect the velocity of money. The equilibrium interest rate. How the Fed influences interest rates. Interest rates and final demand. The monetary feedback. The Keynesian-monetarist debate in detail. Money and inflation. The real rate of interest. Is money neutral?

Chapter 30 The institutional framework of stabilization policy. Potential GNP, the gap, unemployment, and inflation. Costs of falling short of potential. Costs of exceeding potential. Persistent inflation again. Indexing. Direct controls. Strategies of stabilization policy: rules versus discretion.

Chapter 31 Fiscal and monetary policy in action. The federal budget. Built-in stabilizers. Interpreting the federal deficit. The national debt: Who owns it? The burden of the debt. How monetary policy is conducted. Controlling the money supply versus controlling interest rates. Policy coordination and conflict.

Chapter 32 The history of stabilization policy under the Kennedy, John-

son, Nixon, Ford, Carter, and Reagan administrations, with particular emphasis on the relative impacts of fiscal and monetary changes.

Chapter 33 An historical approach to American economic growth. Demographic changes and their effects on growth. Capital accumulation and technical change. Agriculture and industry.

Chapters 34–37 turn outward. They cover the theory and practice of international trade and finance, and the problems of Third World development in a world dominated by developed countries.

Chapter 34 Comparative advantage. The gains from trade and their relationship to demand. The pattern of world trade. Protectionism: tariffs and quotas. The case for protection. Export restrictions. The OPEC cartel. Recent U.S. trade history. Comparative advantage and the pattern of U.S. trade.

Chapter 35 International exchange markets. The balance of payments: current and capital accounts. Determination of exchange rates in the current institutional context. The gold standard and the Bretton Woods system versus the system of floating exchange rates.

Chapter 36 What is economic life like in the Third World? Contrast between the Third and First Worlds. Why is the Third World underdeveloped? Capital, population, education, subsistence agriculture, colonial history. Underdevelopment and international trade. Specialization in primary products. Development strategy. Import substitution versus export promotion. The question of foreign aid.

Throughout much of the world, economics is predominantly Marxist. Chapter 37 presents some of the main elements of the Marxist critique of capitalism. Chapter 38 describes Marxist economic practice in the Soviet Union and China.

Chapter 37 Value and surplus value. The origin of the surplus. Transformation of the surplus into capital. Forces and relations of production. The class structure of capitalism. The reserve army. Capitalist crises. Contradictions and the collapse of capitalism.

Chapter 38 The Russian Revolution. Early failures and successes of socialism. Stalinism, collectivization, and forced industrialization. The structure of the modern Soviet economy. Planning and efficiency. Is Russia socialist? The Chinese Revolution. Reconstruction. The Great Leap. Transformation of the forces of production versus transformation of the relations. The Cultural Revolution. The future of China.

Throughout, there are "boxes" presenting unusual topics, special cases, or extended discussions. Also, each concept is printed in boldface type when it is first presented, and definitions of the concepts are gathered in a glossary at the back of the book.

Teaching aids Each chapter begins and ends with a brief summary of its main points. End materials also include a list of key concepts in the chapter, plus questions for review.

To complement this textbook, there is a set of additional materials: The *Study Guide* (which, along with the test bank, was written by Ann Putallaz with the assistance of Therese Mendola) is designed to help students identify and resolve areas of confusion, and to develop their ability t-

apply theoretical concepts in solving problems. For each chapter, true-false and multiple choice questions and applied problems are presented. The questions focus on concepts with which students frequently have difficulty. Detailed explanations of answers to the questions are provided to ensure that students do not answer a question correctly without understanding why it is correct, and that they are not unduly frustrated by having answered a question incorrectly and not knowing why. The problems help students learn to apply theoretical concepts correctly. Students who work with the study guide will be able to identify sources of confusion, and can build confidence in their ability to apply the material through problem solving. Throughout, an attempt is made to keep students' attention focused on core material, and to encourage them to feel at ease with the subject matter.

The *Test Bank* (available only to instructors) contains multiple choice questions for each chapter. The questions vary considerably in difficulty. Within each chapter, questions are generally arranged sequentially according to the location of the relevant material in the text. Frequently, more than one question is available for a given topic to allow instructors flexibility in designing tests. The test bank is stored in a computer file so that the publisher can provide instructors with individually tailored semester exams. The procedure for ordering these exams is described in the introduction to the *Test Item File.*

The *Instructor's Manual* (prepared by the authors and available only to instructors) is written with the needs of the instructor in mind. It emphasizes the goals of the text, chapter by chapter, and calls the instructor's attention to crucial concepts and diagrams and to areas that students may find particularly difficult. It also gives answers to selected review questions that appear at the end of the text chapters.

A *Transparency Package* containing the most important analytical diagrams is also available from Prentice-Hall.

Acknowledgments

We are deeply indebted to many people for supporting us in shaping this book. For special help from our colleagues at the University of Michigan we want to thank Alan Deardorff, Richard Porter, and Gavin Wright. Our teaching fellows have also given good advice from their classroom experience with the book.

Many scholars at other campuses have provided extensive reviews of early drafts. They include Rich Anderson, Texas A & M University; Marion S. Beaumont, California State University, Long Beach; Peter Bloch, Grinnell College; Daniel S. Christiansen, Albion College; Robert W. Clower, University of California, Los Angeles; J. Ronnie Davis, Western Washington University; James M. Ferguson, Federal Trade Commission; Alfred J. Field, University of North Carolina, Chapel Hill; Max E. Fletcher, University of Idaho; Ralph Gray, DePauw University; John R. Hanson II, Texas A & M University; Barry Hirsch, University of North Carolina, Greensboro; Tom Kniesner, University of North Carolina, Chapel Hill; Kenneth A. Lewis, University of Delaware; Michael Magura, University of Toledo; Robert Moore, Occidental College; Kent W. Olsen, Oklahoma State University; Larry Radecki, Federal Reserve Bank of New York; Michael Salemi, University of North Carolina, Chapel Hill; Allen Sanderson, Princeton University; Len Schifrin, College of William and Mary; James Starkey, University of Rhode Island; John A. Tomashe, California State Univer-

sity, Los Angeles; Holly H. Ulbrich, Clemson University; Tom Ulen, University of Illinois, Urbana; and Jeffrey Wolcowitz, Harvard University.

We have benefitted from research assistance by Barton Lipman, Abdolhamid Mohtadi, George Shepherd, Theodora Shepherd, and Gilbert Skillman.

The publisher has also provided excellent technical support. The editorial gifts and personal commitment of Gerald Lombardi have improved the book on every page. David Hildebrand has steered the book with unfailing talent and devotion. The technical skills, hard work, and sharp eye of Sonia Meyer were invaluable aids to the production of the book.

Among the superb typists who have graced the book are Suzanne Gurney, Judith Jackson, Isabella Leach, Theodora Shepherd, and Joan Susskind.

Finally, our children have sacrificed to make this book possible, by doing without our attention from time to time. We thank them, too, with hopes that they will some day learn from reading it for themselves.

One possible format for a one-semester course

Chapters

1 The Economic Approach
2 Basic Economic Principles
3 Methods and Measurements
4 Demand and Supply
5 Demand and Supply in Action
8 Supply: The Nature of Costs
9 Pricing and Output Under Perfect Competition
10 Monopoly
11 Degrees of Competition
14 Input Markets
17 General Equilibrium
22 An Introduction to Macroeconomic Analysis
23 National Product and Income
24 Equilibrium of the Circular Flow
25 The Multiplier
26 Unemployment and Inflation
27 Financing the Circular Flow
28 Banks and Money Creation
29 Money, Interest, and GNP
30 The Institutions, Goals, and Strategies of Stabilization Policy
31 Fiscal and Monetary Policy
32 Stabilization Policy: The Historical Record

W. H. Locke Anderson
Ann Putallaz
William G. Shepherd

1

The Economic Approach

As you read and study this chapter, you will learn:

- the basic economic issues of scarcity and choice
- the main types of economic goals and economic systems
- the development of economics as a field
- how the economic approach focuses on matters of logic and degree

Economics is the science of production, exchange, and consumption in economic systems. It shows how scarce resources can be used to increase human wealth and welfare.

Its central focus is on *scarcity and choice*. Scarcity is the fundamental economic condition of human life. The resources available to produce goods are limited, so that the goods themselves are scarce. Economic scarcity requires people to make economic choices, and economics is about comparing alternatives and choosing among them.

The need for choices is evident at all levels of life, from personal affairs to matters of worldwide urgency. On a personal level, one might like to have excellent food and clothes, spacious living quarters furnished in style, frequent travel, and so on. Yet because their incomes will provide only modest amounts of these goods, most people must always choose among them. For example, the price of a new coat may equal 50 gallons of gasoline, a weekend trip home, 10 restaurant meals, or a 2 degrees' warmer

room temperature all winter. Each purchase may foreclose buying the others. Such decisions are made routinely by everyone because scarcity requires an endless series of choices.

Companies are also forced by scarcity to make careful choices among alternatives as they convert inputs into outputs. Both a local baker and the huge General Motors Corporation, for example, must decide each day and week how many workers and other inputs to employ, and then use them efficiently in producing bread or automobiles.

At the national level, there are also important economic choices to be made. For example, an increase in the nation's military forces and weaponry might make the country more secure from attack. But the added military spending might have to be obtained by cutting back on programs to inoculate children against disease and to provide medical care to the aged. Better roads may entail worse libraries; more funds for health care may mean less for education. Even more broadly, actions to reduce price inflation may cause national output to fall and unemployment to rise.

To all such small and large choices, economists apply ***economic analysis,*** a system of concepts and logical hypotheses that has been developed over more than two centuries in debates among generations of economists. The debates continue, and economics itself is still changing.

This chapter is your introduction to the field of economics. First you will be introduced to the most basic economic questions. We show how different types of economies handle these questions in quite different ways. Then you will learn the goals by which economic systems can be judged, and why these goals are often hard to achieve.

Then we present the main lines of the United States economy, in which you are naturally involved. The chapter ends by showing the distinctive approach that economists apply to problems. Our hope is that this approach will soon become yours.

Economic systems

Scarcity and choice

Choice and cost are at the heart of economics. Choosing between alternatives, each with its own costs, is the central task of economics. There are three brief questions that summarize the most basic economic problems:

1. ***How much*** **of each good and service should an economy produce?**
2. ***How*** **should these goods be produced?**
3. ***To whom*** **should these goods be distributed?**

These questions involve choice because of a fundamental aspect of human life: *the interaction between scarce resources and the sum of individual wants.* Resources are limited in quantity—when used to produce one good, they cannot be used to produce another.

Given the problem of scarcity, the three questions of *how much* of each good to produce, *how* to produce it, and *to whom* to distribute it become critical. If people wish to have more of one good, they must usually give up something else. As economists often phrase it, a "trade-off" is involved. ***Choosing efficient methods of production is important, because resources are limited: If inputs are used inefficiently, then the total production of goods will be lower.*** Finally, since the economy may not be able to produce abundant goods for all, distribution—dividing the goods among the populace, rich and poor, young and old, and among regions—becomes urgent. Often, the more goods

Table 1 ***The main economic goals***

The Economic Goals	Symptoms of Poor Performance
1. *Efficiency, High Productivity, and Technological Progress*	Low income per person
An efficient allocation of resources	Simple waste, poor management, misallocation among markets
Growth of capacity and output	Low growth rates
New products and techniques	Stagnant products and technology
2. *Fairness in Distribution and Work*	An extreme inequality of wealth and income, not related to productivity or effort
Alternative criteria include: Equality of result and opportunity Rewards for effort and/or talent Meeting people's needs	Unequal opportunity
The sharing of responsibility and of unpleasant work	Unequal access to control and to the best jobs
3. *Stability: Full Employment and Stable Prices*	Sharp recessions, with mass unemployment
Steady jobs for all, suited to workers' skills and preferences	Chronic long-term unemployment
Stable or declining prices	Price inflation
4. *Wider Values*	
Freedom of choice for people as consumers, workers, managers, and investors	Narrow, restricted choices
Security from extreme hardship	Many people in hardship

that one person gets, the fewer goods that will be available for others.

But recognizing that the choices must be made is only the first step. The second involves making these choices by deciding which trade-offs are best. To deal with these choices, every nation has evolved its own peculiar economic system, whether by historical chance or by design.

Each economy is a system in which the production and distribution of goods are organized around people's wants. No ***economic system*** is tight and rigid, like a factory assembly line, nor as loose as, say, the world's weather "system." It is more like a large city, in which many people live, work, and play. All of them go their own ways, and yet their actions mesh with and respond to one another.

Economic goals

There are many criteria for appraising the performance of economic systems. Table 1 summarizes the main ***economic goals*** that economists agree on. Certain goals that are relatively easy to measure, such as output, incomes, or unemployment rates, can be reduced to a few simple numerical scales. Though the measures are not perfectly precise, they are widely used to compare different economies and to study long-term economic trends.

Even if every detail on the list of economic goals and indexes were accepted by all economists, judging an economy's performance would still be difficult for several reasons. First, the goals range from those that are fairly easy to measure, such as total output or employment rates, to those

that are extremely hard to quantify or measure, like freedom. Second, people may differ about which goals are most important. For example, you may think that freedom of choice is paramount, but the person sitting next to you may give the most value to efficiency, while a third person may rank fairness or security first.

There is still another difficulty in applying the goals and indexes. It is a fundamental problem, which arises from the difference between positive and normative economics.

Positive and normative economics

Economic analysis operates on two levels: positive and normative. ***Positive statements*** are about facts, *what is* or has happened, or how certain conditions are related to each other. "Six percent of the work force was unemployed last year" is a positive statement, which could ultimately be affirmed or rejected by scientific measurement.

Normative statements are about value judgments, about *what ought to be.* A normative statement usually expresses ethical standards and values. People naturally accept those normative statements that fit their own values and reject those that do not. Thus, "Six percent unemployment is too high" is normative, comparing the fact of 6 percent unemployment with a standard of what is unreasonable.

It is hard to find a widely acceptable *normative* standard of reasonable performance. Judging economic performance takes several steps, each one difficult and debatable. First, you must measure what is happening and why (positive knowledge). Then you develop your normative standards of good performance, such as high incomes and low unemployment. This step includes assigning values to the relative importance of the goals. Finally, you compare the actual performance with the standards.

The possibilities for normative disagreements are great. One economist may rate high incomes as most important, while another may emphasize avoiding unemployment. With the goals being given different values, judgments about how well an economy is doing may differ sharply. Nonetheless, economists often must make such judgments as best they can. Indeed, judgments of both positive and normative elements are a delicate but routine task for economists from the moment they begin introductory economics.

Thus far we have presented economic systems in general terms. You also need to grasp some details of the U.S. economy in order to understand the economic issues discussed in this book. At this point, the U.S. economy will be summarized to give you a factual background. Later you will be able to fill in the outline with much richer detail and insights.

The main features of the modern U.S. economy are quite similar to those of other industrialized economies, such as Japan, Canada, and Western Europe. As Figure 1 suggests, these economies are composed of four major sectors. First, there are the basic *utility* industries (power, transport, communications, city services) that underlie and support all other economic activities. Next are the *primary* industries, such as agriculture, fuels, and mining. Most of their goods flow into the *industrial* sector, whose activities range from the processing of raw materials to the manufacture of goods. Finally, there is the *service* sector, which includes sales, repair, government, education, and health. As Figure 1 indicates, these four sectors of the economy are interrelated. All sectors also draw on financial and labor markets. Moreover, a set of *governments*—federal, state, and local—absorbs large amounts of output and influences activities in many markets.

Table 2 shows these sectors' relative importance. Initially, agriculture was

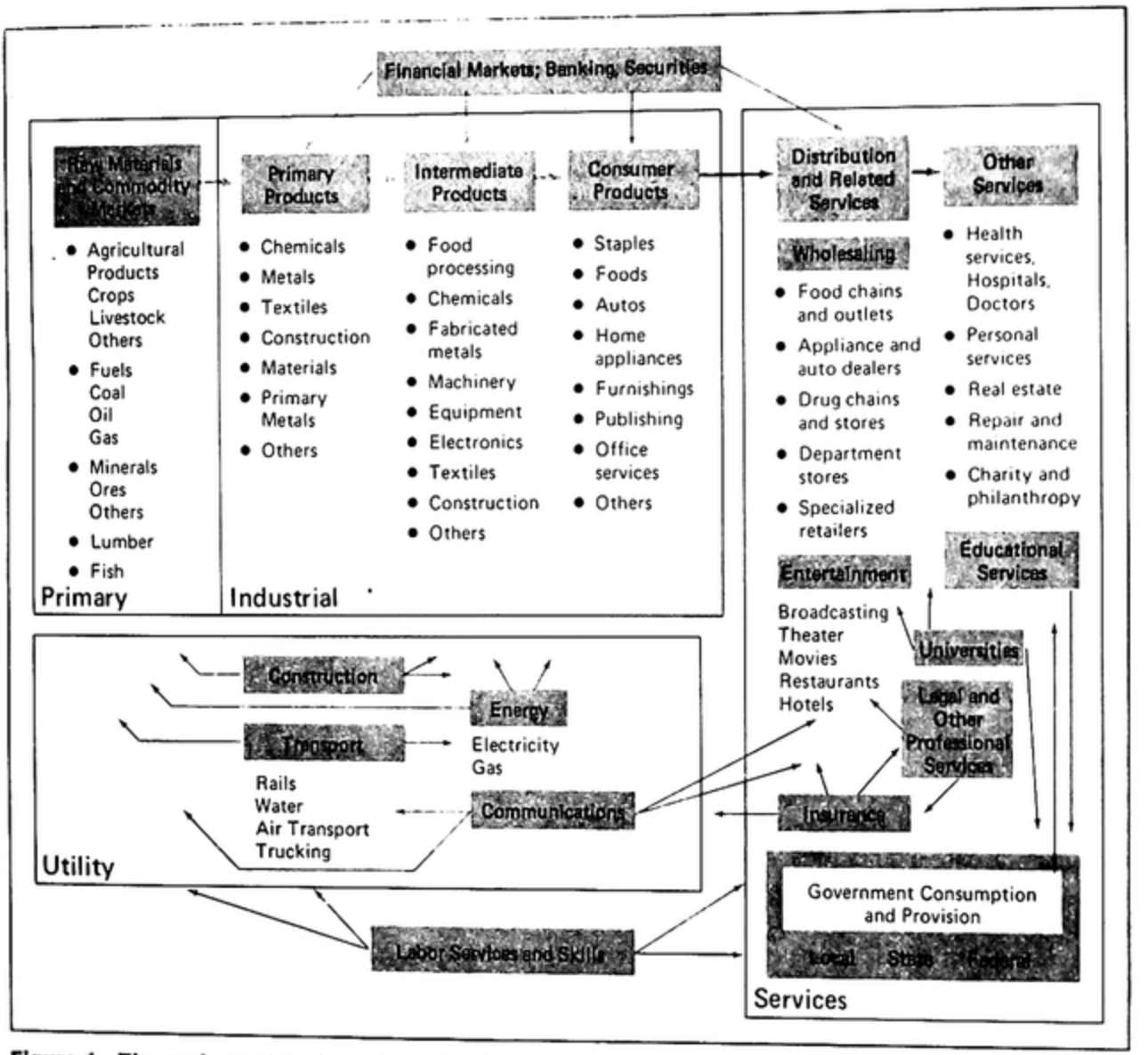


Figure 1 The main sectors in an industrialized economy

The three main parts are utilities, industrial, and services. Each part has many industries and thousands of individual markets within it. Although this diagram only sketches the complex flows and arrangements among them, it does suggest the main kinds of sectors where microeconomic activity occurs in the economy.

dominant; most people worked on farms. Manufacturing and utilities became more important as industry grew from 1870 to 1920. Today, the service sectors (including government services) have come to dominate the mature U.S. economy. Especially since 1930, governments have absorbed a growing share of national production.

The U.S. economy has an enormous capacity to produce goods and services. Part of this capacity comes from the country's rich *natural resources*. Another part comes from the *skills* that Americans bring to their work. The remaining capacity comes from human-made *capital*, embodied in factories, offices, roads, rails and harbors, cities, farms, houses and apartments, electric and telephone systems, and all the rest.

The growth of production occurs partly because these factors increase, raising the capacity of the economy. Techno-

Table 2 The main sectors of the U.S. economy, 1840–1950

Share of Total Value of Output (percent)				
	1840	1900	1960	1980
Primary				
Agriculture, forestry, and fisheries	48	23	4	3
Mining and construction	7	10	7	6
Manufacturing (Foods, clothing, wood products, chemicals, oil, metals, machinery, transportation, equipment, and others)	12	24	30	26
Utilities (Transportation, communications, electric, gas, urban services)	6	11	8	8
Services (Wholesale and retail, financial, real estate, and others)	23	25	38	41
Governments	3	6	12	15
Total	100	100	100	100

Sources: *Statistical Abstract of the United States* and *Historical Statistics of the United States*.
Note: Details do not add to 100 because of rounding.

logical progress also makes the factors more productive. New methods of production are embodied in new investments, using the latest techniques. The populace also grows more skilled, more able to operate complex processes.

Population and production correlate closely with the distribution of income. High incomes are focused in industrial areas and in states with large farms. As the map shows, these high-income areas are mainly in the Northeast and parts of the Great Plains and West Coast states. Low incomes are concentrated in small farms and rural areas, particularly in the South and Southeast.

This diversity of agriculture and industry is matched by a great *diversity in the size of enterprises*, ranging from tiny one-person shops to the American Telephone and Telegraph Company (AT&T: the Bell System), which employs one million people. You may picture the U.S. economy as made up of corporate giants like AT&T and General Motors. Yet, most economic acitivity occurs in small and medium-sized firms with 1,000 or fewer employees.

During economic *recessions*, which seem to occur every four or five years, several million people are thrown out of work, some of them for long intervals. Among some groups of workers, such as young black men, unemployment stays above 30 percent for long periods. Unemployment rates this high are not simply an "interesting" economic phenomenon: They are also a severe social problem.

The production of goods and services creates income and wealth. **In the United States, as in most developed countries, the *distribution of this wealth* is distinctly unequal.**

One feature of this inequality is that while the United States has one of the highest per capita incomes in the world, *poverty* is still widespread. In 1980, about 25 million people, or approximately 12 percent of the U.S. population, had incomes below the officially designated "poverty level" of about $8,000 per year, unable

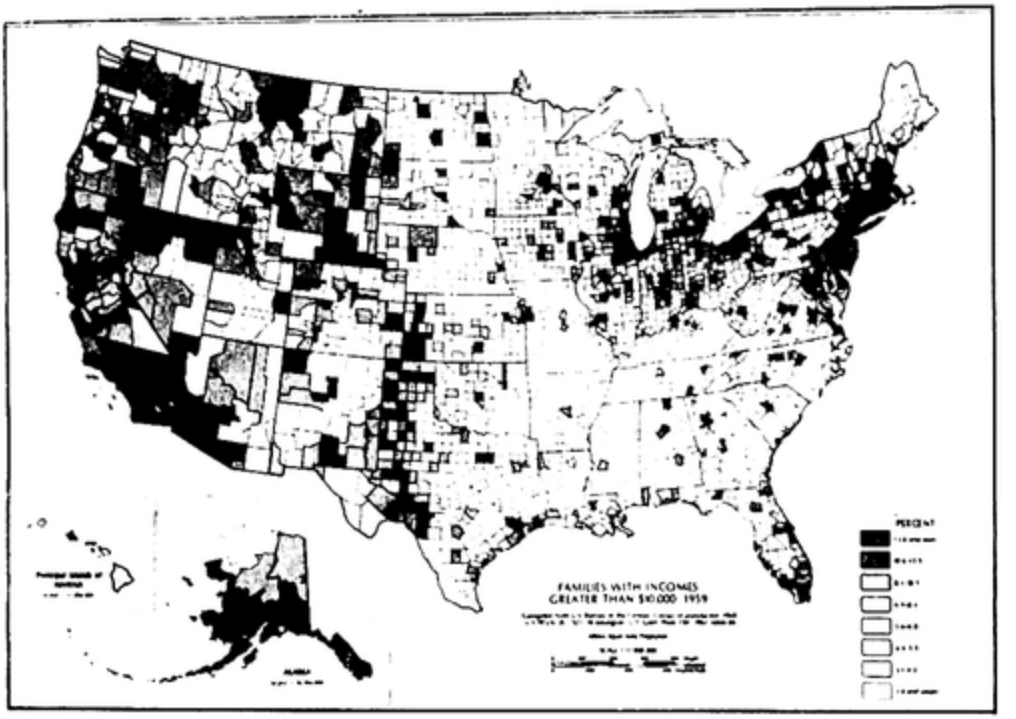


Figure 2 Geographic distribution of family income
The darker shadings indicate counties with higher incomes.
Source: U.S. Bureau of the Census

to afford more than the barest necessities of life.*

In short, the U.S. economy mobilizes a vast array of capital, raw materials, and labor to produce a staggering $3,300 billion worth of goods and services per year. Yet, impressive as the performance of the U.S. economy is, its functioning is far from perfect. Its economic record, like that of all the countries of the world, is both good and bad. It resembles a glass that is partly full of water—or partly empty, depending on your viewpoint. While most people in the United States are rich by world standards, severe economic problems remain.

One of the important skills you will develop by studying economics is the ability to look clearly at the U.S. and other economies and to assess their strengths and weaknesses. This process of analyzing problems and searching for solutions is the key to understanding how economic thought has developed. As you will see in the next section, much of the history of economic thought has been shaped by economic problems themselves. New questions have forced economists to search for new answers and techniques.

*The poverty benchmark varies among regions and towns to reflect differing costs (in rents, heating, etc.). The benchmark values have risen over time as general price levels have increased.

Economic analysis

Early economic thought

In ancient Sumeria, Babylonia, China, Egypt, Greece, and Rome, people tried to understand and improve the economic process. Plato and Aristotle analyzed trade, production, and values at some length. All of the modern concerns—production, jobs, prices, monopoly power over

markets, public works, money, interest rates—were vigorously debated more than 2,000 years ago. Much of the analysis was primitive and vague, but the concern about economic issues was there, along with some important insights.

Economic thought, which had become increasingly sophisticated under the Roman Empire, was then forgotten or neglected during the "Dark Ages" from A.D. 400 to 900. It was only in the 1200s, with St. Thomas Aquinas's discourses on the "just price" and interest rates, that economic analysis was revived. From 1400 to 1700, expanding trade, the discovery of new worlds, and the rise of early industry caused economic development to move ahead faster than economic thought.

From 1600 to 1750, the dominant economic doctrine was mercantilism. Mercantilists believed that economic wealth was embodied mainly in precious metals, whose possession by the state helped to enlarge military capacity and national power. To amass this wealth, states relied on taxes and trade restrictions. The French Physiocrats (1760–1780) argued against the mercantile view. They stressed that wealth is economic capacity rather than gold, productive resources rather than money.

Classical economics

Classical economics differed from most of the earlier economic schools in its insistence that national wealth is the capacity to produce goods or material products. Yet, unlike the French Physiocrats, the classical economists viewed capacity as resources, labor, and capital, a far broader definition than that of the Physiocrats.

Adam Smith's *Wealth of Nations*, published in 1776, synthesized for the first time the ideas of Smith's immediate predecessors into one grand system of analysis. Smith (1723–1790) stressed that people were guided by the "invisible hand" of self-interest in their economic choices. These free choices would lead to the development of exactly the appropriate mix of skills and capital necessary to increase national output or "opulence."

Smith's optimism was attacked by later economists. Thomas Malthus (1766–1834) argued in 1798 in his *Essay on the Principle of Population* that unchecked population growth would strain the world's food supply and pull the population back toward poverty. David Ricardo (1772–1823) analyzed how economic growth could overcome the scarcity of resources, but he also foresaw barriers to economic growth. Together, Malthus and Ricardo earned economics the reputation of being "the dismal science." Yet, Smith's optimistic outlook seemed justified to many, as British industry boomed and spectacularly increased its output.

Some exponents carried the doctrines of free choice to extremes. Defenders of unbridled capitalism argued that even the ugliest industrial exploitation was inevitable and acceptable to achieve the efficiency of free markets. These writers were called the "Manchester School" after their location in that leading British industrial city. It was the excesses of early capitalism with its riches for the few and grinding labor and poverty for the many that stirred the French Socialists, along with Karl Marx in *Capital*, to stress the cruel and self-destructive nature of capitalism.

Neoclassical economics

By the late 19th century, Europe had become so wealthy that many Europeans could afford to purchase goods far beyond the subsistence level. This greater latitude for consumer choice was one reason that several brilliant economists began working independently in the 1860s and 1870s on the theory of consumer choice: Stanley Jevons (1835–1882) and Alfred Marshall

Five Leading Economists

Each of these five famous economists represents a different stage and emphasis in the development of economics.

Adam Smith (1723–1790) was professor of moral philosophy at the University of Glasgow. He resigned to travel and write *The Wealth of Nations* (1776) and then remained an influential economist while an official in the British customs service. A wry-humored Scot and a close colleague of the controversial philosopher David Hume, Smith was both learned and worldly wise. Though a leading advocate of free choice in private markets, he recognized that private enterprises are often inefficient and prone to fix prices with their supposed competitors.

David Ricardo (1772–1823) was a successful London stockbroker and a member of Parliament, as well as the leading classical economist of his time. Gifted in abstract thinking, he submitted the leading economic problems of the times to penetrating analysis. He attacked restrictions on agricultural prices and other market barriers. He originated the concept of economic rent as received by owners of land and other resources, developed the theory of relative prices (giving labor a main role), and created the basic analysis of international trade. A cheerful, kindly gentleman, he was widely loved and respected.

Karl Marx (1818–1883) was first a German revolutionary (co-author of the *Communist Manifesto*, 1848), and then a powerful social historian and economic analyst in his three-volume *Capital*. In attacking capitalism, he attributed value mainly to labor rather than capital (thus taking Ricardo's view to the extreme), and forecast more suffering for workers until a revolution replaced capitalism with socialism. Eccentric, irascible, and usually poor, he toiled in London on his immense and often barely intelligible volumes. At his death in 1883, only one had been published. The whole set has had enormous







influence, and Marxian economics is officially accepted by the Soviet bloc nations and China.

Alfred Marshall (1842–1924) was an early pioneer of neoclassical economics. As the Cambridge University professor of political economy, Marshall was highly influential in the fledgling field of economics and as a commentator on public affairs. Though skilled in mathematics, he always subordinated mathematical technique to the content of economic ideas. His *Principles of Economics* (1890) is still a powerful and lively summary of neoclassical economics.

John Maynard Keynes (1883–1946) was brilliant in economics, successful in commodity and stock speculation, a leading intellectual in public debates, and a lover of high culture. He was a member of the Bloomsbury Group that included Virginia Woolf and Lytton Strachey, and started on a diplomatic career. After spectacularly (and correctly) denouncing the peace treaty after World War I, he became a leading expert on British monetary affairs. The calamitous Great Depression of the 1930s stimulated his *General Theory*, which explained why an economic collapse might not be self-correcting. This made him instantly the leading exponent of modern macroeconomics. Witty, generous, and tireless, just before his death he helped devise the world monetary system that made possible the long prosperity of 1945–1970.





(1842–1924) in England, Carl Menger (1840–1921) in Austria, and Leon Walras (1834–1910) in Switzerland were the leading neoclassical pioneers. **The framework of neoclassical economics erected at that time still dominates Western economic thought.**

Leon Walras showed in 1874 how an economy's many markets fit together to form a complete economic structure. He stressed that each market both influenced and was influenced by every other market, as part of an internally consistent whole. A century later, Walras is still viewed as one of the greatest mathematical economists.

By 1890, Alfred Marshall had combined in his *Principles of Economics* much of the best of the older classical analysis and the newer neoclassical analysis. Marshall's text analyzed cost, productivity, demand, and output choices in an economy composed of competitive markets. These topics are in the branch of economics called *microeconomics*, which focuses on

one economic actor (such as a consumer or producer) or one market at a time. In England and Austria, theories of capital and business cycles were being developed during 1870–1920, stressing the self-correcting tendency of competitive markets. This belief in the innate stability of national economies was shaken in the United States by the financial crash of 1929 and the Great Depression of the 1930s.

In 1936, John Maynard Keynes (pronounced "canes") published the *General Theory of Employment, Interest and Money*, in which he held that depressions may not be self-correcting. Instead, deliberate government policy might be needed to move the economy of a mature Western nation back to full employment. Keynes' work led to the modern field of *macroeconomics*, which analyzes the economy as an aggregate. Keynesian economists who stressed the need for government spending to cure depressions had developed macroeconomics to full flower by the 1960s.

The Great Depression of the 1930s severely embarrassed economists, for their traditional theories could not explain or cure the devastating economic stagnation. Keynes' systematic analysis of the causes and cures of depression, plus the economic recovery late in the decade, restored much of their self-confidence. World War II reversed conditions by causing an all-out boom in production. Economists were drawn into managing much of the price controls, planning, and financing of that war.

They then helped guide Europe's postwar recovery and America's long boom during 1945–1965. The steady economic growth of the 1960–1966 period in the United States was a triumph for Keynesian economics. Leading macroeconomists were at the president's elbow, "fine-tuning" the economy to achieve full employment with little inflation.

But that brief golden era was the peak of the Keynesian success. President Johnson escalated the Vietnam War; its great expenditures overstrained the economy's capacity and started rapid inflation. The steep rise in oil prices after 1973 spurred the inflationary rise, and a high and persistent rate of unemployment accompanied it. Keynesian analysis had few clear answers to the problem of simultaneous unemployment and inflation.

New situations call for new theories and tools, and the search for new answers began. Monetarism, "supply side" economics, "Reaganomics," and other approaches have been advanced since the mid-1970s. Macroeconomics remains unsettled, and the problems remain unsolved.

Microeconomics has also encountered urgent new problems. For example, the sharp rise in energy prices since 1973 has forced major new choices about the use of coal, oil, nuclear power, and solar energy. Race and sex discrimination in employment are stubborn problems that have an important economic component. Does the minimum wage law hurt or help minority workers? Should government's roles in fields such as education, railroads, and communication be reduced? If so, in what form and to what degree? Should governments set safety standards in factories? If so, which methods of protection are most economical? Should antitrust policies reduce corporate power? These and the many other microeconomic issues are controversial, often stirring the sharpest political debate. Yet, microeconomists insist that rational answers to them begin with a careful weighing of the economic costs and benefits.

Economists have always worked on both theory and practical problems, often or even usually immersed in intense controversies. Before 1930, they were a small band of generalists, mostly teachers on college campuses, who did some research and occasionally debated public policies. They were often brilliant, but were few in num-

ber. Now the field consists of a sizable army, some 50,000 specialists, most of them with advanced degrees. A few economists are still generalists. But most—whether they teach, do full-time research, advise political leaders, work in government or for private industry—concentrate in a specialized field and focus on narrow technical problems.

The literature

Like other scientists, economists analyze issues, conduct research, and publish their findings. *Their writings form the literature of economics:* articles, books, and reports in which concepts and facts are debated. New ideas are advanced to replace old ones, old ones are defended, and the sifting process retains—it is hoped—the best ones. Often older ideas are "rediscovered" and replace some of the newer ideas! Economists write with a purpose: to change ideas and to make their own reputations. The literature embodies those debates as the field evolves.

*The core of the literature consists of the professional journals and specialized books, written by professional economists.** The literature consists of layers, as illustrated in Figure 3. New professional articles and books form the core of the literature. Written by economists to persuade and inform their fellow economists, these core publications provide a growing stream of new ideas and facts. Textbooks (like the one you are reading) are rarely at this creative core of the literature, although they do reflect professional standards of objectivity.

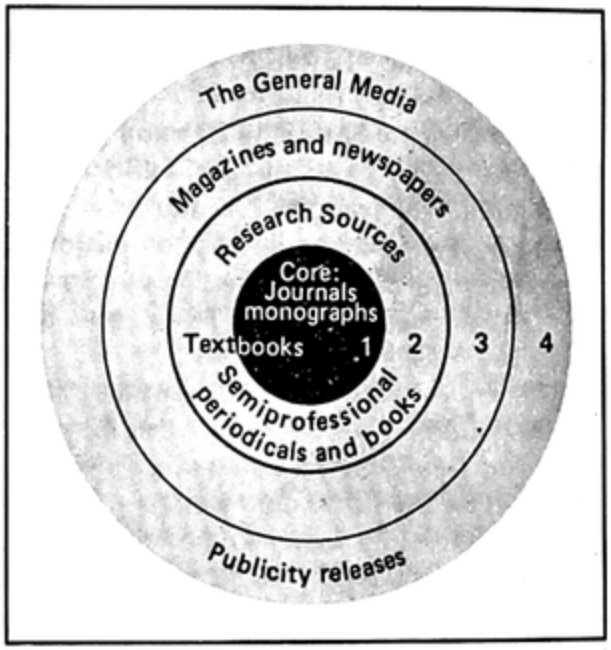


Figure 3 Layers in economic literature

Texts try to present the main concepts that have come to be accepted among economists, not to change economic thinking or engage in debate. Magazine articles and many government reports are far from the core: They are often merely second-hand—and one-sided—accounts of what economists are saying.

The field of economics is always evolving; it never reaches a final set of tools or answers. By 1860, classical economics seemed to be complete and refined; then came the neoclassical revolution. By 1965, Keynesian economics seemed to have the answers, but new problems requiring new answers arose in the next ten years. Throughout economics, the debates go on.

Reading in the literature can give you a sense of its changes and the ability to judge each piece for its value in the ongoing debates. Each piece of professional work at the core helps to make the changes. Some books—like Smith's *Wealth of Nations*, Marx's *Capital*, and Keynes' *General Theory*—have a massive effect. The thousands of others move ideas

*The leading U.S. journals are the *American Economic Review*, the *Quarterly Journal of Economics*, and the *Journal of Political Economy;* there are perhaps six other important ones. There is also a host of prestigious specialized journals, such as the *Review of Economics and Statistics*, the *Journal of Economic Theory*, and the *Journal of Economic History*.

by inches: some forward, some perhaps backward! You can enjoy economics for such elements of change and human drama, while mastering the basic tools.

Schools and groups

Every field has its "schools" and groups. They have differences of opinions that may go to the very roots of methods and values. The groupings often involve subtle shades of opinion along a spectrum, rather than diametric clashes. Still, it is often helpful to recognize these points of view.

One is composed of "liberal" economists, trained in both neoclassical and Keynesian analysis from the 1930s to the 1970s. They expect the market system to perform reasonably well. When it fails, they often propose corrective government action: to prevent or ease recessions, reduce pollution, regulate industry, promote competition, and the like.

To the right are conservative economists—often called classical liberals, or "free-market" economists, or the "Chicago School." They rely on people's private choices as the rational basis for the economic system. Therefore, they trust the free-market system to deliver efficiency and technological progress. They deeply distrust government actions, which interfere with the incentives applied in a market system.

On the left, "radical economists" including Marxists and other critics see fatal flaws in free-market capitalism. In their view, the capitalist system is dominated by financial power; it exploits workers and is subject to frequent depressions and mass unemployment. Despite a patchwork of cures designed by "liberal" economists, the radicals believe that the system is basically unsound and unfair to most of the people. The system needs to be changed outright, they say, not just improved or allowed to run free.

The economic approach: Matters of logic and degree

Economists have a distinctive approach to problems. It is best to present that ***economic approach*** now, so that you can develop it in your own thinking from the start. It has several main parts.

Logic and matters of degree *Economic concepts are matters of logic,* which can be refined and analyzed to a high degree of precision. Used clearly, these concepts can cut razor-sharp through complicated human affairs, exposing the essentials. *But economic conditions are also matters of degree.* The logic needs to be applied with a sound sense of reality, weighing the size of conditions and the force of many crosscurrents. Thus, logic and matters of degree are both involved. Simple formulas cannot solve difficult problems. If they could, problems would no longer be tough, and economics could be done by clerks applying rules of logic. Economics is partly a science, with clear logical lines, but it is also an art, requiring good judgment and good sense.

Systems and deeper causes Above all, economists see the economy as a system with interacting parts. These parts—consumer, factory, or market—can be studied separately, but their actions also affect one another. To be able to see the whole system and its parts—and to trace changes as they ripple through the system—is the mark of economic skill and wisdom. Thinking about systems is second nature to economists, for economics itself is a system of related concepts.

So each issue is tackled with an eye to its extra economic elements. Economists always cast their net wider than most immediate issues, searching for hidden, unexpected costs that others have neglected. They probe for real causes, ignoring the superficial events that crowd the daily news.

What is *really* causing high unemployment, or inefficient factories, or poverty, they ask? What have others forgotten to analyze? Again and again, economists show how a narrow problem has larger roots and effects.

Comparing effort with reward (costs with benefits) Economists are always comparing alternatives. Usually they compare an effort with a reward—a cost with the benefit it gives, a sacrifice with its resulting return—and they ask, "Is the reward worth the effort?" Thus, if a firm adds $6 million to its costs by building a new factory, will the resulting increase in its sales revenue more than equal the added costs? If unemployment rises from 6 to 8 million, will inflation drop by 4 percent or 2 percent or not at all? If $10 million is spent to control pollution, will the added value of the cleaner air justify the cost? In each case, is the value of the change *worth* its costs? In numberless practical cases, economists ignore Shakespeare's advice that "Comparisons are odious." Comparisons are, in fact, the very stuff of economics.

Tracing alternatives—their costs and benefits—is the economist's standard work. Often, though not always, it is easy to figure out the facts, to see what *is;* but the economist must also study what *might be.* That helps to understand present conditions more clearly.

Marginal changes The economist expects things to change by degrees, not by opposites. A little more here, or a little less there, will cause a degree of change in other things. These small or "marginal" changes can often be compared accurately: ***Such marginal analysis is crucial throughout much of economics.*** If one talks only of either-or's, of jumps from one condition to another, one is probably not talking good economics.

Clinical analysis Most economists care deeply about poverty and injustice, progress and freedom. Their aim is to help people by improving the economy's performance. But the analysis must first be objective, to show correctly the exact processes at work. Only if the processes are seen clearly can the causes and results be judged accurately. Then, and only then, does the careful economist apply that knowledge to show which remedies might cause the economy to do better.

Common sense All valid economic concepts can be tested in the end by common sense. Indeed, you will sharpen and extend your own good sense as you become an economist. You will need to master concepts clearly and use them with technical skill, as you would any new language. But that technical precision needs to be allied with good sense and balance.

An independent mind Above all, economics requires independence of mind. Chapter by chapter, you will learn increasingly to *think for yourself.* Economic analysis can give sharply diverging answers, even though the core concepts are pretty much agreed upon. You will need to be skeptical of every answer and claim, skeptical even of the concepts themselves. Once they are tested and familiar, the economic tools and habits of mind will become part of your thinking.

Summary

The aim in this chapter has been to show the essentials of economics as a field of study. The leading points of the chapter are summarized below.

1. Scarcity is a fundamental condition of life and of economics. Since there

are not enough resources to produce outputs sufficient to satisfy all wants, hard choices must be made.

2. Economic systems must cope with the choices set by the three most basic economic questions: *how much* of each good to produce, *how* to produce the goods, and *to whom* the goods should be distributed.
3. In judging economic performance, and in any other economic analysis, one needs to separate *positive* statements concerning what exists, from *normative* statements concerning what ought to be. Positive matters are strictly factual, though often hard to grasp; normative issues involve ethical values.
4. Economic issues have been debated intensely for many centuries, but modern economics is generally agreed to have begun with the publication of Adam Smith's *Wealth of Nations* in 1776. The classical economists (including Smith) viewed wealth as productive capacity. They stressed the contributions of free choice and specialization to the growth of national wealth. The advent of neoclassical theory after 1870 and Keynesian economics in the 1930s provided many of the main tools of modern economic analysis.
5. There are three main schools of economic thinkers:
   a. "Liberal" economists, who believe that although the market usually gives good results, governments can often correct market failures by specific actions, rules, and reforms.
   b. Conservative or "classical liberal" economists, who trust the market system thoroughly and distrust all government actions.
   c. "Radical" economists, including Marxists, who see basic flaws in free-market capitalism and urge that the system be radically reformed or replaced.
6. It is important to learn the economic approach in applying economic concepts to the conditions of the world. The main elements of the economic approach are:
   a. To deal both with matters of logic and matters of degree.
   b. To see the economy as a system.
   c. To compare alternative costs and benefits.
   d. To compare small or "marginal" changes, not extreme jumps from one condition to another.
   e. To apply an objective analysis.
   f. To keep an independent mind.

Key concepts

Economic analysis

Economic system

Market economy

Centrally planned economy

Economic goals

Positive and normative statements

Economic approach

Questions for review

1. "The problem of scarcity, which haunts so much of the world's population, can hardly be said to exist in the United States. With its vast array of re-

sources, the U.S. economy can produce almost unlimited amounts." Discuss.

2. Using Table 1 as a starting point, rank in order of importance the five goals that you think are the most significant in judging an economy's performance.
3. Explain whether the following statements are positive or normative:
   a. "The unemployment rate was 6.8 percent in 1982."
   b. "American capitalism is an efficient system of production."
   c. "The availability of consumer goods is an important criterion of economic performance."

2

Basic Economic Principles

As you read and study this chapter, you will learn:

- the circular nature of flows in the economic system
- microeconomic concepts including opportunity cost, marginal choices, equilibrium, and specialization by comparative advantage
- macroeconomic concepts including inputs and outputs, aggregate demand and income, and policies to stabilize economic activity

Even the most complicated card games are played by applying a few simple principles. Take bridge, for example. The play of the hand by the declarer, or person who has the bid, presents an endless variety of problems. There are more different arrangements of the cards than you probably care to think about. Yet a winning strategy is always pieced together by combining a few elementary stratagems: the finesse, ruff, drop, squeeze, strip, holdup, and throw-in. What distinguishes the master from the ordinary player is the imagination, concentration, and sense of timing with which he or she applies these principles. The principles themselves can be learned from a book. Skill in applying them comes from experience in playing the game. If you have ever played bridge seriously, you know how exciting it is when you first realize that you can rise to the challenge of a new problem, piecing together a strategy of your own.

Economics, too, has a relatively small number of principles. Just as in bridge, the master is distinguished by the imagination, concentration, and sense of timing with which he or she applies these few principles. Gaining this skill is simply a matter of practice. And again, success is exciting.

In Chapter 1 we introduced the economy as a system. Our first task in this chapter is to present that system's main parts: *households*, which consume, and *enterprises* (also called firms), which produce. For simplicity at the start, we will focus on these two fundamental parts. Two other sectors are important but will be saved for later chapters. One is *government*, which influences the economy through its spending and taxing, and through regulations that influence the behavior of households and firms. The other is *foreign trade*, which includes what an economy buys from and sells to other countries.

First we will discuss the economy's basic circular flow, to show how the parts are related. Then we will define the choices and motives of households and enterprises. Next we will introduce the concept of the ***market***, the arena where households and firms meet. Finally, we will unite these economic units—households, firms, and markets—in a more complete version of the **circular flow of goods and money.**

The economy as a system

Every economy is a system in which the production of many goods is organized to fit people's many wants. The system has an underlying circular pattern, which connects its many markets and economic actors. As illustrated in Figure 1, the two main kinds of economic units in a system like the U.S. economy, *households* and *enterprises*, are linked by a circular pattern of economic activity.

The choices and actions of these two main units are the driving force of economic activity. In their households, people make two sets of decisions: (1) *selling the inputs they own*, primarily their labor, but also land and capital, and (2) *buying goods with their incomes*. Meanwhile, enterprises *engage in production*, using the labor and other inputs bought from households. The goods produced by the firms are sold ultimately to the households. In Figure 1, the inputs and goods flow clockwise, as shown by blue arrows. The money paid for these items flows counterclockwise.

The interactions of households and firms bring together the two sides of economics: demand and supply. The action occurs in two sets of markets: that for inputs and that for outputs. In the input markets, households offer their labor, land, and capital. Firms buy those inputs at prices set in the markets. Notice that

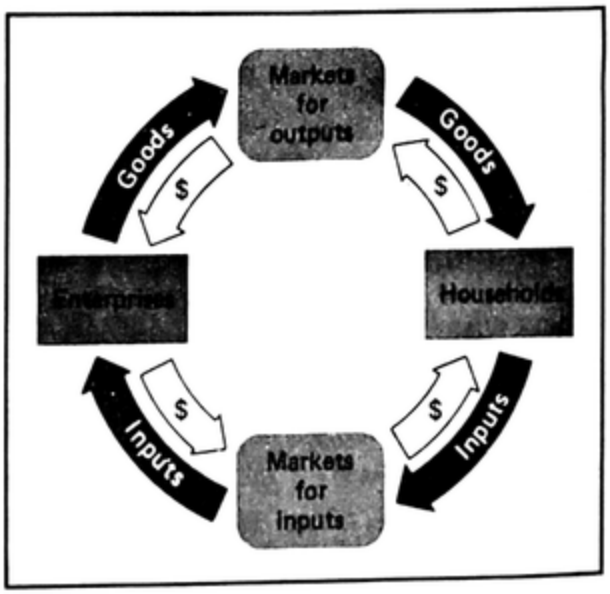


Figure 1 The simple circular flow of goods and money

Inputs flow from households to input markets, are exchanged there, and flow on to enterprises. Production occurs in the enterprises, the resulting goods flow to output markets, where they are sold, and then flow to households to be consumed. The physical items flow clockwise, while money flows in the counterclockwise direction.

households are the suppliers and firms are the consumers in input markets.

The two roles are reversed in output markets. There the households are the consumers while the firms are the suppliers. The households' demand (how much they want to buy at alternative prices) interacts with the firms' supply (how much they will offer at alternative prices). The result is a definite price and quantity in each market.

The whole circular system, therefore, yields specific levels of production and consumption in each period. It answers the three basic questions—**what** goods to produce, **how** to produce them, and to **whom** to distribute them. Demand and supply interact throughout. For example, if in 1983 there are 3.1 billion bushels of wheat, 1.2 million new houses, and 6 million new bicycles sold in the United States, those amounts reflect demand and supply working throughout the system.

Households: The decision units for selling inputs and buying goods

A ***household*** is any unit in which people live and make decisions about work, consumption, and the disposal of personal property. In America, the conventional household has been the nuclear family. But many households consist of only one person or temporary pairings of some kind. College students sharing an apartment, for example, would be considered a household. Most households have one or more members who work, but some are composed of retired or unemployed couples or of singles who consume but get no income from work. Despite this variety, all households must make the same basic decisions.

In making *consumption decisions*, households budget their spending and assign it among goods to buy. Their incomes come mainly from work, though some households receive income from property or live on pensions or welfare. Households must keep their purchases within their incomes, often making hard choices within small budgets. Households can go into debt, of course, but the debts must be repaid eventually. Income can often be raised by working longer hours, perhaps by having more household members work, or by taking on extra part-time jobs. But those ways of enlarging income have their limits, since they cut into family life and leisure time, which are themselves goods.

The ultimate determinants of how much people are willing to work in exchange for income, and what they will buy with the income, are *human wants*. Everyone wants or needs food, clothing, and shelter. These needs are basic, universal, and easy to understand. Nonetheless, much of the world's population has to struggle to gain even a minimum of these rudiments.

In general, *economists take personal preferences as a given*. Preferences, in fact, make up a person's identity and are accumulated over a lifetime of experience. They need not be rigid, requiring fixed amounts of each good for each person. On the contrary, preferences involve *degrees of choice* among goods, as prices vary. In applying their preferences, people continually adjust their choices as conditions change.

Preferences also govern people's choices about the other half of household choice: *choosing work and managing personal property*. The work choice is often complex. Which kind of work to do, for which employer, at what rate of pay, and under what hours and other conditions—these must all be decided, often day after day and month after month. Choices about training for work must also be made, in deciding about college or other specialized programs. Often these choices involve several household members at the same time, in a complicated balancing of interests.

Managing the household's property also expresses preferences. Assets can be

held in a variety of forms, with varying risks and possible profits. Stocks, bonds, and other paper assets; farms, houses, and other real estate; jewelry and other valuables—all these offer numberless choices to households with money to invest. The choices can be an important expression of preferences.

Finally, households themselves *also engage in productive activity:* washing clothes, preparing meals, health care, cleaning, do-it-yourself projects, child care, and other work. Households are minifactories that have consumption and production activities. Unlike firms, however, their production is not for sale on markets.

Enterprises and inputs

Enterprises (or firms) are the basic units of production, converting inputs into outputs. *Like households, enterprises face constraints. They must repeatedly make choices, keeping their expenditures within their money income.* They obtain that income by selling their outputs at specific prices to their customers. They spend to acquire the inputs necessary to produce the finished products. The difference between a firm's income and its production costs is *profit.* ***Maximizing profit is the primary aim of private enterprises.*** To maximize profit, firms must choose the right level of output to produce, buy the right mix of inputs, and keep production as efficient as possible.

Along with the private firms that dominate most market economies, there are also public firms, cooperatives, and nonprofit enterprises. Examples include most hospitals, city bus systems, and your own college. They do not seek profits, at least not as a formal objective. But they all have to balance income and expenditures.

Recall that *households are the basis of consumer demand:* Incomes, market prices, and preferences shape the demand for goods by households. Similarly, *enterprises are the basis for supply:* The potential income from sales, prices, and the productivity of inputs shapes the supply of goods. ***Inputs*** are resources, which are purchased by enterprises, processed, and turned into outputs. They are not only an important part of the flow of goods and services—they also are the basis of the economy's ability to produce.

Categories of Inputs For centuries, economists have spoken of ***labor, land, and capital*** as the three fundamental categories of inputs. These are the ***factors of production,*** the endowment of resources that makes production possible and also determines how large it can be.

The inputs are brought together in production, as workers use equipment in workplaces to process materials into goods. The combinations of inputs vary greatly from industry to industry. In some industries, a few workers may manage enormous machines, cooking up large volumes of complex chemicals. In other industries, masses of workers apply simple hand tools to wood, metal, or cloth. In still other industries, workers may use pens, typewriters, computers, or other apparatus—along with their mental capacities—to do office work. In all cases, though, output is achieved by some combination of the three major classes of input.

Labor is the physical and mental effort of people applied during periods of work: hours, weeks, years. The economy's stock of skills ranges from simple labor to highly trained technical and professional capacity. Since training itself is expensive and time consuming, the stock of human talent is a form of human capital that is an important part of the economy's productive capacity. Yet even highly trained specialists do not work in a vacuum. They need the appropriate capital equipment with which to work.

Capital is the set of productive resources—buildings, machinery, roads,

and other tangible means of production—that has been created by production and investment in the past. Capital is distinctive because (1) it is made by human production rather than by "nature"; and (2) it is used to produce other goods or services. Capital embraces a great variety of equipment: Buildings, roads, harbors, sewers, dams, electric wires, and other engineering works are all examples of capital. A violin and a broom are also pieces of capital; so are a sledgehammer, a jumbo jet, a typewriter, a courthouse, and a mine.

A primitive society may have vast natural resources but little capital other than knives, hammers, and hoes. An industrial economy invariably has a large stock of capital, much of it highly complex and specialized. The classic image of capital is the "dark satanic mill" of a century or two ago, a vast factory or arsenal with huge furnaces where people toiled among heavy machinery. Compared to that, much capital today is both quiet and complex.

All economies use capital in production. If the capital is owned privately, the economic system is called ***capitalism.*** The U.S. economy is mainly a capitalist system because most of its enterprises are privately owned. A successful firm enlarges its sales, holds its costs down, and reaps a financial surplus of sales revenue over costs. *This surplus or profit goes to the firm's owners.* The workers are paid for their labor, while the owners of capital get the extra value that arises when the business prospers. Of course, the owners also bear the risk of possible losses.

Land is the broad term for both (1) geographic area and (2) natural resources. Since most production activity occupies space, land is inevitably a factor of production. Farming uses large amounts of land; office work requires very little. The price paid for the use of land varies, in some cases to great extremes, according to the land's location and quality. Thus a rich acre in Iowa or a plot in lower Manhattan Island draws a much higher purchase price or level of rent than does scrubland in Utah or swampland in Alabama. In cities, the location of each parcel of land is crucial in determining its value, while for farming the quality of land is usually critical.

There are many types of natural resources. They occur in land, streams, oceans, and air. The main categories of natural resources are:

Nonrenewable: fuels (coal, oil, gas), topsoil, ores, chemical deposits, natural beauty sites.

Replaceable at great cost: wilderness, certain rivers and lakes, clean shorelines.

Renewable: other rivers and lakes, forests, fish, grass cover.

Virtually inexhaustible: fresh air, solar heat.

Now that you are more thoroughly acquainted with the household and business sectors, one fact should be obvious: Both sectors of the economy are continually confronted with choices. Households and firms must accept some alternative plans of action and reject others, always mindful of the constraints under which they operate. To choose the "right" course of action, they must have some goal—some principle—by which they can organize their production and consumption decisions. What principle motivates them? The economist's answer is *self-interest,* or in more technical terms, *maximizing:* each unit doing the best that it can for itself.

Maximizing behavior

Economists consider the primary motive of households and firms to be self-interest. The technical term for it is ***maximizing.*** Households select the group of purchases that will *maximize their satisfaction within*

the limits of their incomes. On the supply side, firms choose the combination of inputs and level of output that will *maximize their profits.* Self-interest is the proper basis for microeconomic analysis because almost all people and firms are guided by it almost all the time. Your family, the corner grocer, General Motors Corporation—these and the millions of other units all pursue their own self-interest.

Maximizing organizes people's actions, just as a string passed through a row of beads converts them into a necklace. Each purchase by a family is part of its whole effort to do its best, given the constraints of its income. Each such purchase involves a choice among alternatives in which more of one thing means less of another. More bread means less potatoes, and for most parents, putting a child through college means less dining out and fewer vacation trips. This balancing among choices to achieve the most satisfaction is how consumers maximize.

For firms, too, maximizing gives coherence or pattern to a string of specific actions. The owners of a local restaurant, for example, adjust their buying among meat, flour, plates, cooks' and waiters' time, and set the prices of the meals, all as steps along the road to making money. Maximizing their profits is what spurs and unifies all of those actions. It is the same, too, with the other 15 million private firms in the United States, plus millions more in other countries.

The *"As If"* proposition This pursuit of self-interest need not be complete, conscious, or infallibly correct in every detail. Real choices often involve trial and error, but they are still organized by the principle of maximum advantage. For example, consider a choice that seems to violate the principle: you buy a car that does not work. The result is a loss of satisfaction, not a gain. Yet you were trying to maximize: The "lemon" is merely a factual error, not a lapse of motive. Such errors are part of the inevitable variations and adjustments that occur even for the most finicky maximizers. For economic analysis, what matters is that maximizing is the consistent underlying motive.

All maximizing theory assumes is that people's choices are *mainly* guided by self-interest most of the time. They will therefore act *as if* they were maximizing their advantage and follow the patterns predicted by maximizing behavior. This *as if* proposition is simple and fundamental, and it gives the logic of maximizing a firm practical basis.

Market exchange

A market is a grouping of the buyers and sellers of a good, in which they make exchanges. Their choices and exchanges determine the good's price and quantity. An economy is a mosaic of many such markets. Each market is an arena in which the supply of a good and the demand for it meet. Their interaction determines the value of the good, as shown by its price.

Market exchange In any exchange, two parties agree to a mutual trade. In a market exchange, one person trades a physical item (a *good,* such as a shirt, or a *service,* such as a taxi ride or a piano lesson), while in exchange the other person trades money to purchase the item. The amount of money exchanged per unit is its *price.*

The exchange is voluntary. Each of the parties regards it as better than no exchange at all. That crucial fact deserves a careful explanation: Both sides in an exchange might hope for better terms than they actually get. The seller would prefer to obtain the highest price possible, while the buyer would like to pay as low a price as possible. Yet some price in the middle may be both higher than the seller's alter-

native opportunities and lower than what the buyer would have to pay in the next-best choice. Therefore, the sale at that price gives some advantage to each side.

Value and price As these exchanges occur, the value of the good, reflected in its price, is determined by supply and demand in the market. ***The market value is the market price.*** The price will move up or down as changes occur on the demand or supply side. For example, a house near your campus may be priced now at $60,000. If more people suddenly wanted to buy housing near campus, the value or price of that house would probably increase. In a properly functioning market, the price adjusts until the buyers (on the demand side) and the sellers (on the supply side) want to exchange the same amounts. ***"The market clears" as demand and supply are brought into line: There is no physical shortage or surplus of the good. The price, then, reflects the workings of both supply and demand.***

Economics focuses much of its analysis on markets. That is natural because much of economics since Smith's and Ricardo's time has been the theory of value, of what determines prices, and markets are where economic value is mainly determined.

You have looked at households, enterprises, and the process of exchange in markets, all in some detail. It is now time to fit these pieces together, treating the economic system as a whole.

The circular flow

Figure 2 shows the economy's underlying circular pattern in more detail than Figure 1. Guided by the primary motive of self-interest, the 70 million households and 15 million enterprises go about their daily business of purchasing, consuming, and producing. What happens in one part of the system, in one household or one enterprise, can affect conditions throughout the system.

As you can see from the diagram, this interaction of households and firms occurs in three broad types of markets. First are the *markets for inputs,* where households supply labor, resources, and the services of their financial capital, accumulated from past savings. Firms are the buyers. They pay wages, rents, and capital incomes, such as interest and dividends. Second are the *markets for consumer goods.* Third are the *financial markets* in which households supply new financial capital from their current saving out of income. In exchange, firms promise to pay interest and dividends in the future. These promises may be in the form of stocks, bonds, or savings accounts. The firms use this new capital to finance their purchases of machinery, buildings, and other inputs used in their processes of production.

The whole process seems simple, and is, just as the circulatory system of the human body is basically simple. Inputs of labor, land, and capital services flow through the factor markets from households to firms. In firms, they are combined with raw materials and semifinished goods to produce outputs. Some of these outputs flow to other firms and some to households. The flows of money payments run in the opposite direction. ***The circular flow is thus really two separate but equal flows—the "real" flow of goods and services through countless markets, matched by a "money" flow of dollars moving in the opposite direction.***

Simple as it is, the process unites an overwhelmingly complex array of units and markets. All manner of consumer choices, from bland to bizarre, are serviced by millions of firms. Some markets are tiny local ones; others, dominated by big business, are national in scope. Yet the circular flow operates in and connects them all.

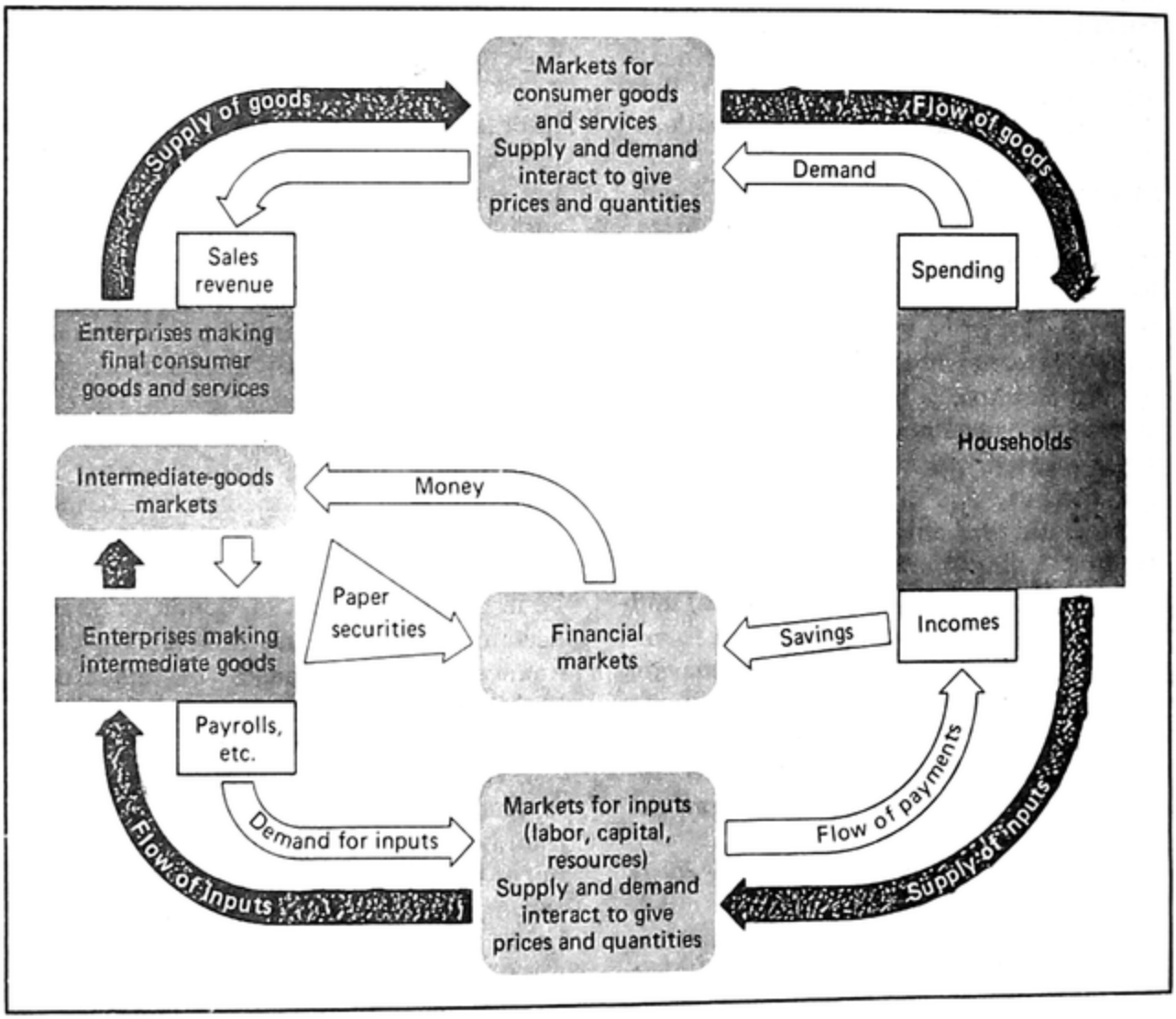


Figure 2 The detailed circular system

Households and enterprises make the main economic choices, and deal with one another in markets. Firms sell their products on output markets. Some of their output goes to other firms, and some goes to households. Households get their income from selling labor, land, and capital services to firms in factor markets. Part of this income goes back to enterprises to pay for consumption goods, and part goes through the financial markets, as households exchange their current saving for promises of future income. The circular flow of goods, services, and money permeates the entire economy, as millions of households and firms deal in millions of markets.

Microeconomic principles

Having studied the main parts of an economic system—the households, enterprises, and markets—you understand that these parts fit together to form a system. Economists who deal with the individual sectors or markets of this economic system work in the branch of economics known as *microeconomics*. ***Microeconomics*** concentrates on the smaller details of the economy, on parts of the whole. Studies concentrating on household or firm behavior, or on supply and demand conditions in individual markets, all deal with microeconomic issues. They often analyze only a tiny part of the economy at a time.

In fact, the three basic questions—*what* and *how much* of the various goods

to produce, *how* to produce them, and to *whom* to distribute the goods—are all microeconomic issues, since each question focuses attention on one part of the circular flow. How much to produce focuses on household wants, how to produce focuses on firm behavior, and for whom to produce focuses on distribution of income among the households in the economy.

The principles of microeconomics are few, but they are powerful and of general validity: opportunity cost, marginal conditions, diminishing marginal effect, scarcity and efficiency, comparative advantage, equilibrium, and public choice. The more familiar you become with these principles, the better you will be at microeconomic analysis.

Opportunity cost

Opportunity cost is a fundamental concept of microeconomics. *It is the value of the best alternative.* This basis for measuring true cost pervades all economic choices. Every choice involves taking one action rather than others. For each choice, *the person compares the benefits and costs of each alternative, trying to get the maximum net benefit (the benefit minus the cost).*

The *benefit* side is usually clear and easy to appraise. The satisfaction and advantages from acquiring a car, meal, concert, or even a year of college are usually matters of direct personal experience. But the *cost* side is often much more subtle and difficult to measure. The true economic cost—the opportunity cost—*may* be simply the price you pay for the good. But instead, the opportunity cost may differ sharply from the simple dollar figure because other real costs are incurred.

These other real costs are often hidden, or indirect, or ***implicit costs,*** to use the technical term. *An implicit cost is a sacrifice that does not involve the paying of money.* An example: Arthur goes to a movie the night before an examination for which he has not prepared. The movie's price is $3.50, which is the direct cost. But by not spending that time studying, Arthur gets a C rather than an A on the examination. That lower grade is the implicit cost of the movie. The total opportunity cost of the evening at the movie is $3.50 plus the difference between an A and a C.

Four examples of opportunity costs are given in the box on the next page. In all cases, opportunity cost depends on the true value of the best alternative. That value often requires a careful appraisal for hidden elements. For example, the true cost of college noted in the box does include the job earnings sacrificed by not spending the year at a paying job. But there are further elements. Only the earnings *after taxes* are true costs, since the taxes would not have been kept by the worker. And the cost of living while a student is *not* part of the true cost of college because it would have been spent in either case.

Marginal conditions

Most economic choices involve ***marginal*** changes. To be most precise, marginal means adding just one more unit. ***A decision "at the margin" compares the benefits and costs of small changes.*** For a consumer, the choice involves buying one more unit of a good, such as a hamburger or a gallon of gas. Producers' marginal choices involve using one more unit of an input or producing one more unit of output.

Marginal choices have been the central focus of neoclassical microeconomics since its inception over a century ago. The approach is valid because *most maximizing behavior by consumers and producers does indeed involve marginal choices.* Most decisions involve small adjustments rather than radical changes. Alfred Marshall

Calculations of Opportunity Cost

Commonsense Instances

1. With just 2 days before exams, you could either study economics exclusively, thereby raising your course grade from a B to an A, or only chemistry, raising that grade from a C to an A. The *opportunity cost* of an A in chemistry is therefore getting a B in economics rather than an A.
2. Your parents bought a house for $20,000 some years back. They could sell it for $45,000 now. The accounting cost of their staying in the house now is $20,000. The *opportunity cost* is $45,000.
3. College costs for you this year are $3,000 tuition and $3,000 living costs (food, lodgings, clothing, etc.). That totals $6,000. But you could have earned $8,000 after taxes on a full-time job if you weren't in college. The *opportunity cost* is $3,000 tuition *plus* the $8,000 not earned at a paying job: This equals $11,000 (the $3,000 for living costs would have been spent in either case). The opportunity cost differs from accounting costs both in amount and in the kinds of items included.

An Example of a Commercial Decision

4. A firm has $1 million of retained earnings left over from the previous year. The firm can invest the funds in either Plan A or Plan B. Plan A will pay 15 percent rate of return, while Plan B offers 18 percent. The accounting cost of the firm's use of the funds is zero. The *opportunity cost* of the funds for any use other than Plan B is 18 percent.

stressed this by adopting the Latin phrase *Natura non facit saltum*—nature does not make jumps—as the motto on the title page of his *Principles of Economics*. Marginal analysis is firmly rooted in reality.

Moreover, the effects of marginal changes can be precisely defined and measured. Therefore, marginal analysis has developed many exact conclusions, which can be verified by practical tests.

Six main marginal concepts are at the heart of modern economics. They are listed in Table 1. Though few in number, they are among the most frequently used economic tools. Marginal decisions are the cutting edges of economic activity. They determine value and cost in markets. They are the standard for judging the efficiency of the economy.

Diminishing marginal effect

Marginal concepts guide economic choices made by consumers, enterprises, workers, investors, and the rest, in all parts of the economy. They give definite outcomes because of an important marginal principle: ***diminishing marginal effect.*** *The effect of any good or input tapers off as more of it is used.* This holds true for both demand and supply situations. It is also a common phenomenon that can be verified by personal experience.

Table 1 The main marginal concepts

1. *Marginal utility:* The change in satisfaction gained from consuming one more unit of a good.
2. *Marginal product:* The change in output arising from the addition of one more unit of an input, assuming that other inputs are held constant.
3. *Marginal cost:* The change in cost resulting from the production of one more unit of output.
4. *Marginal revenue:* The change in revenue resulting from the sale of one more unit of output.
5. *Marginal benefit:* the economic gains (in utility, satisfaction, or other values) obtained from having one more unit of a public or private good.
6. *Marginal propensity to spend:* The proportion of a change in income that will be spent.

First, consider *demand*. As you consume more of any good, the pleasure from each added bit diminishes. The day's first glass of orange juice may seem wonderful; the fifth glass is less refreshing; the fifteenth would make you sick. You prove diminishing marginal effect every time you pour a certain amount of juice to drink, but not an ounce more. Why did you not have a fifth or twelfth glass? Diminishing marginal effect is the answer. This eventual decline in marginal pleasure or satisfaction (called "marginal utility") from consuming additional units of a good is called the *law of diminishing marginal utility*.

Next, consider *supply*. As more of an input is added to production with another input held constant, its contribution to output—called its "marginal product"—declines. For example, the first waiter hired for a small restaurant accomplishes a lot; the fifth waiter will add less because four waiters are already at work, while the sixth or seventh waiter might just get in the way of the others, perhaps subtracting from production rather than adding to it. This effect is called the *law of diminishing returns*. It applies throughout factories and shops, and helps to determine how much of each input to use.

Diminishing marginal effect sets the range within which each type of action (consuming a good, using an input to produce an output, or investing in capital) makes sense. It sets limits and discourages going "too far." Everyone consumes scores of food items, not just a lot of one or two. Firms use many inputs together, not just one. Because of diminishing marginal effect, economic choices reach balances among many elements, rather than being lopsided or extreme.

Efficiency and scarcity: The production-possibility boundary

You have read about how people decide between alternative courses of action (opportunity cost), how marginal conditions are crucial, and how choices are limited by diminishing marginal effect. This continual choosing or deciding among alternatives is necessary because of a fundamental human condition: *scarcity*. No person—no country, for that matter—can have as much of everything as he or she wants. Given scarcity, it is important not only to produce the most "desirable" goods, to make the right output decisions, but also to produce these goods *efficiently*. Efficient production conserves society's scarce resources.

Efficiency is important because human wants outrun what can be produced from the available resources. If the resources can be used efficiently, then the maximum value of production can be obtained. Efficiency has meaning and importance because of scarcity.

Efficiency is achieved when a given level of output is produced using the least amount of inputs. Any switch of inputs to other uses will reduce the total value of output. Efficient production is also referred to as the *least-cost* method of production.

Firms seek to minimize their costs by adjusting their production so as to use rel-

atively cheap inputs. If an input grows scarcer, its value (as shown by its price) will rise. Firms will respond by using less of this input, substituting others for it at the margin. As all firms respond in this way to relative scarcity, the whole economy will adjust so as to minimize costs.

The concepts of efficiency and scarcity are clearly shown by the simple diagram in Figure 3, called a ***production-possibility boundary.*** The figure illustrates not only efficiency and scarcity, but also choice, opportunity cost, and diminishing marginal effect. The principles shown by this diagram, which represents an economy that *produces two goods,* can be applied equally well to an economy producing the usual vast array of goods. (The vast array of goods, however, cannot be easily portrayed in a two-dimensional diagram.)

The economy represented by Figure 3 allocates its resources between two categories of goods: food and shelter. The curve itself depicts the supply side of the economy. *It represents the maximum combinations of food and shelter that can be produced, given the country's resources and technology.* Each point on the curve is therefore a combination of the two goods that can be attained when all resources are being used, and used efficiently. Each point is an alternative to other points that could be reached by using the same available resources. Thus, point F is for 480,000,000 units of food and 20,000,000 units of housing, while point C is for 150,000,000 food units and 110,000,000 housing units. All other possible combinations are also shown on the curve.

The enclosed shaded area represents the economy's capacity to produce. That capacity may be large or small, so that the economy may be wealthy or poor. Curve 1 in Figure 3 represents high capacity, compared to the limited potential output shown by Curve 2.

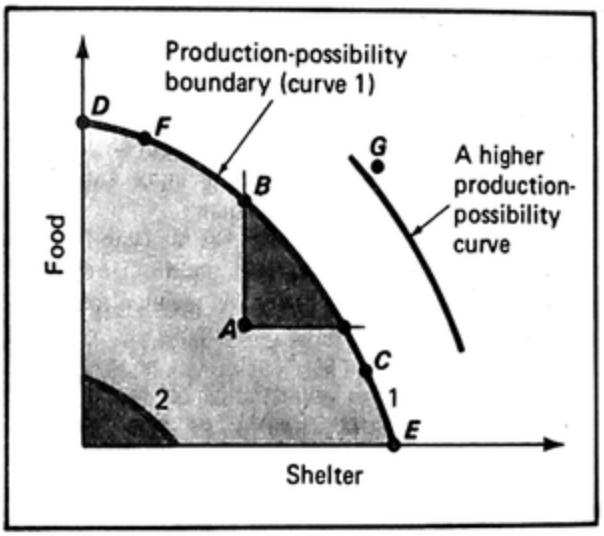


Figure 3 A production-possibility curve for food and shelter

Combinations of output that fall in the shaded zone can be produced. Amounts outside the zone are beyond the economy's capacity. The boundary of the zone is the production-possibility curve (curve 1). Point A is an inefficient point, for more of both goods could be produced by moving to the northeast. Curve 2 is a production-possibility curve for a smaller economy with fewer resources, which can produce much less.

The rounded shape of both curves reflects diminishing marginal returns. Thus, to get more food beyond point B (up to point D), farmers have to resort to increasingly barren land. Also, they have to hire construction workers (skilled only at building houses) as farm workers. The conversion grows harder and harder to do successfully. At point D, the last carpenter—who is least competent at farming—has finally been put to work in the field.

The physical amounts of food and shelter shown are:

	Food	Shelter
A. (Inefficient)	200,000,000	60,000,000
B.	400,000,000	60,000,000
C.	150,000,000	110,000,000
D.	500,000,000	0
E.	0	125,000,000
F.	480,000,000	20,000,000
G. (Impossible)	450,000,000	110,000,000

The boundary itself separates the attainable from the unattainable quantities of goods. All points on or *inside* the production-possibility boundary are possible. All points *outside* the boundary, like Point G, are unattainable. They are beyond the economy's capacity. To operate *inside* the

production-possibility curve, at a point like A, is inefficient. If the economy is at Point A, it means that some resources are unemployed and/or that resources are being used inefficiently. More of both goods could be produced by moving closer to the boundary, as the economy achieves a more complete and efficient use of its resources.

The curve has a negative slope, downward from right to left, reflecting *scarcity*. To get more housing, the populace must give up some food, since resources must be taken from one good in order to provide more of the other. Scarcity does not apply inside the boundary, for from any inner point more of both goods can be obtained.

Once production has moved out to a point on the boundary, then scarcity requires that hard choices be made. ***Each marginal increase in the amount of one good involves a marginal sacrifice or decrease in the amount of the other good.*** Therefore, the production of each good has an ***opportunity cost,*** which is shown precisely by the slope of the curve. The population can have more food only if it is willing to give up some shelter, and vice versa. If the economy must give up three units of housing to obtain an additional unit of food, then three units of housing is the opportunity cost of that additional unit of food. Inside the curve, of course, the opportunity cost of an additional unit of food would be zero, since more of both goods could be had without sacrifice.

Diminishing marginal effect or returns is reflected in the rounded or concave shape of the production-possibility boundary. It shows that shifts of resources to either good will encounter diminishing marginal returns. To get larger and larger amounts of shelter, for example, one has to draw away resources that are best suited to growing crops and put them to work on increasingly crowded building sites. This is illustrated in Figure 4. Therefore, there is diminishing marginal productivity—as units of input produce successively smaller additions to output—in producing shelter. Thus, the rise in housing production tapers off, as it takes more and more inputs (and sharper decreases in the amount of food) to produce another unit of housing.

For example, at Point A in Figure 4, if the economy gives up 50 million units of food, it can add 25 million units of housing. But at Point B, a 50-million-unit reduction in food will add only 5 million units of housing. Thus, the opportunity cost of each good rises as more and more of it is produced. The same effect also works in the other direction. As the economy produces increasing amounts of food, the additional units of food are gained only by larger and larger sacrifices of housing.

The only escape from the scarcity along the curve is to shift the whole boundary outward by expanding the economy's capacity to produce. The expansion can come from several main sources, illustrated in Figure 5. *Population may grow,* enlarging the labor force, as in Panel I. Rapid population growth could move the boundary out swiftly, but of course the additional people would also add to consumption, so that the production *per person* might not rise.

Technology might improve, as in Panel II, giving higher production from unchanged inputs. Many economists credit such technological innovation with a large share of actual economic growth in the last century. Technology need not enlarge the capacity for both goods equally. For example, rapid improvements in farming technology (in equipment, fertilizer, and hybrid seeds) could shift out the boundary as shown by the dashed line in Panel II. Indeed, such a contrast between progress in farming and in housing construction probably has occurred in recent decades.

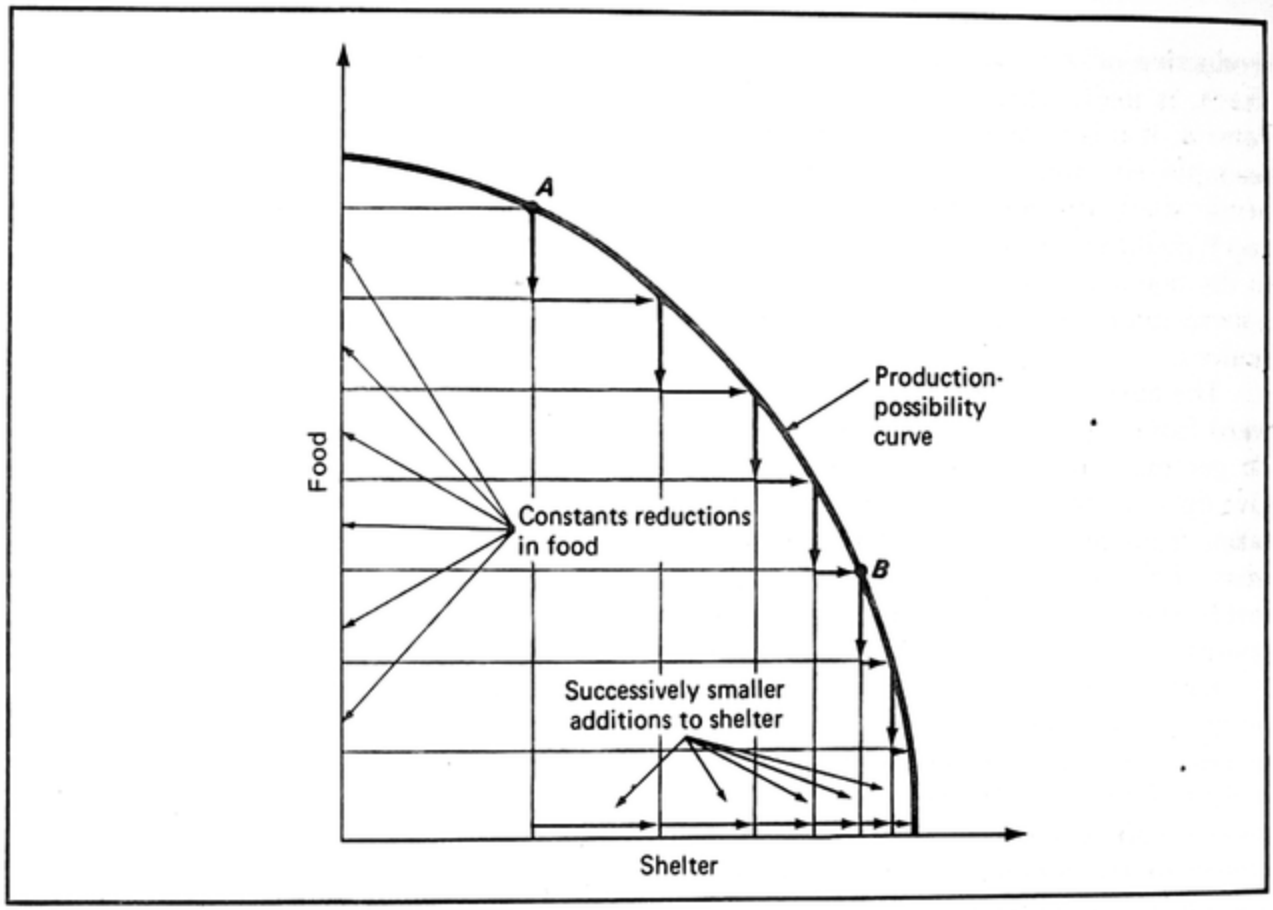


Figure 4 Diminishing marginal effect

Along the production-possibility curve there is diminishing marginal effect in both directions. One direction is shown by the stair-step arrows. They indicate the amount of additional shelter that can be produced with the resources released by giving up constant marginal amounts of food. In each step, the same amount of housing is sacrificed, but the transferred inputs give smaller and smaller additions to shelter.

Growth in the capital stock will also shift out the boundary. But such growth requires investment, which diverts production from present consumption. Panels III and IV illustrate this choice. The two goods are now assumed to be consumption goods and investment goods. Higher amounts of investment goods in one period cause the boundary to move out farther in the next period. In Panel III, year 1 involves a level of investment represented by point A. The curve for year 2 is therefore well outside the initial curve. But in Panel IV, a lower amount of investment occurs, represented by point B, so that the boundary moves out very little by year 2.

The amount of resources devoted to investment is of great importance. Japan is often said to be like Panel III because its investment levels are above 20 percent of total production and its growth rates average about 10 percent annually. In the U.S. economy, investment has been about 10 percent of total output, and the growth rate has been only 5 percent per year. Continued over many years, these trends may bring dramatic changes in economic affluence and power.

Finally, we return to the simple food-shelter choices in Figure 3 to note that *not all points on the curve are equally desirable.* At Point D the population might eat well but have no protection from the weather or other dangers. At Point E, in contrast, the population might live in stately mansions but would soon starve to death. A de-

gree of balance is obviously needed, somewhere in the middle of the range. But whether the choice will lean toward food (as at Point F) or shelter (as at Point C) would depend upon consumer preferences, on the demand side, rather than on supply conditions.

Equilibrium

An* equilibrium *is a condition reached when all influences balance each other out, so that there is no pressure for further change. An economic equilibrium may exist for an individual consumer or firm, for a market, or for the entire economic system. Much of economics is about defining such equilibrium situations and the forces that may disturb them.

An equilibrium is usually not a state of rest, with action coming to a halt (such as three balls in a round bowl). Rather, it combines various forces in a way that keeps the resulting outcome (such as price or quantity) the same. The concept is common in fields outside economics (such as chemistry), as well as in ordinary life. For example, a moving bicyclist is in equilibrium; so are a sleeping person and an airplane flying at its cruising speed. There is motion, perhaps with sharply counter-

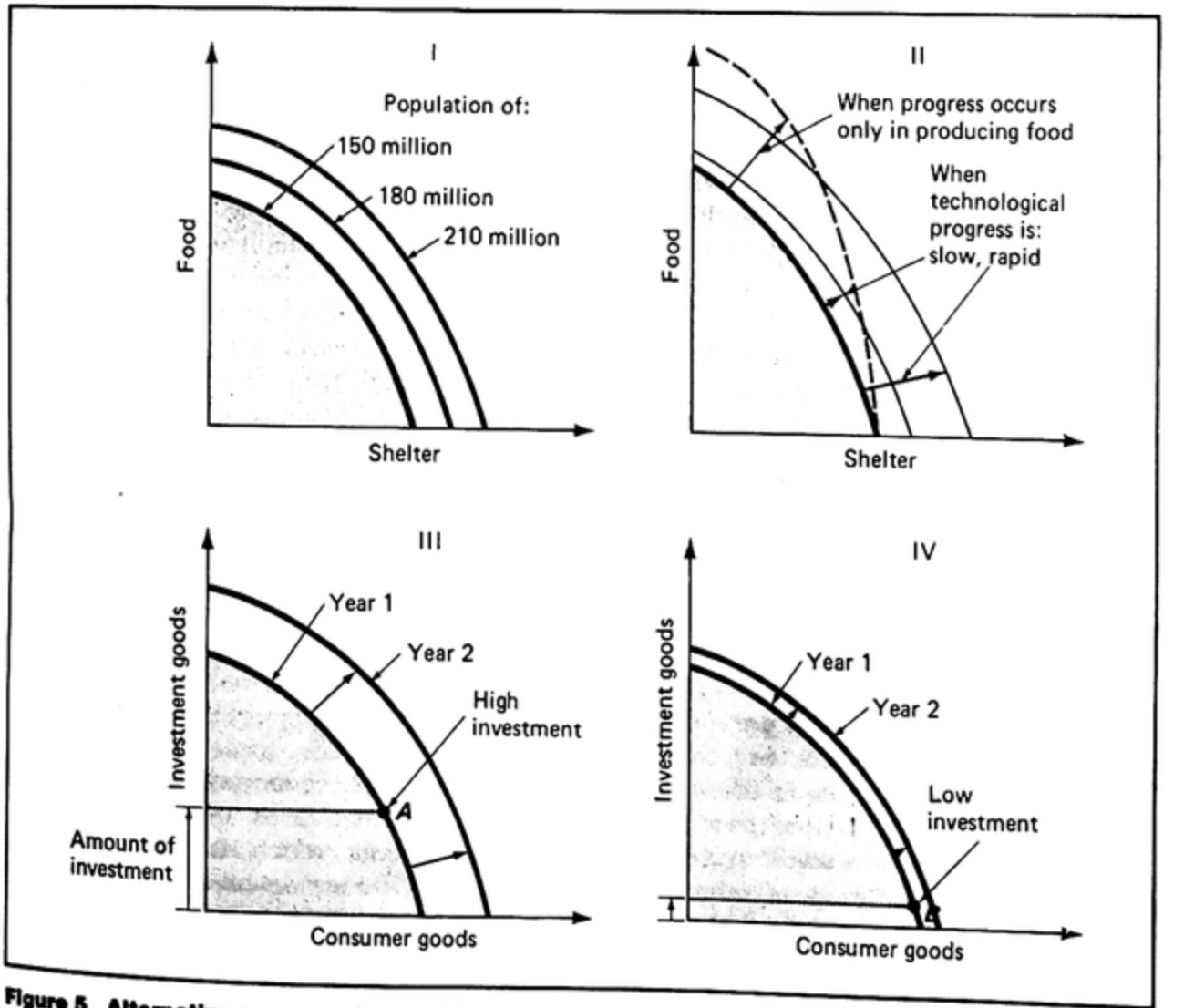


Figure 5 Alternative sources of growth in production possibilities

posed influences (as in the terrific forces within the airplane's jet engines), but *the result is an unchanging level of activity.*

In economic equilibrium, people are satisfied to maintain their level of activity, so that the economic outcomes—in prices, quantities, or any other item—are constant. The outcomes may be fast or slow, high or low, normal or bizarre, but they remain constant.

Such pure situations are unusual, though they precisely convey the concept. Actual economic processes involve continual changes, which require adjustments. Therefore, a total, exact equilibrium with perfectly constant values is rare. *Yet most people behave as if they were aware of what their equilibrium choices would be and are moving toward them.* Most markets are in the range of their equilibrium outcomes most of the time.

The economic system as a whole also continually moves toward an equilibrium. In this case, the balance must be between total demand for all goods and services and total or aggregate supply. If people want to purchase more goods than are available, firms will see their inventories going down and orders up, and realize that output should increase. If the units in the economy do not want to purchase all of the goods and services available, then firms will see their inventories piling up and orders falling, and they will decrease output. It is only when total demand and supply balance that the economy is in equilibrium.

In equilibrium, not only are total supply and demand balanced, but each and every market must also be in equilibrium, for any disturbance in one part of the economy will affect the whole system. For example, a bumper crop of wheat will have effects on other food markets, which will spread to restaurants, grocery stores, and so on. As these ripples spread out, they dwindle—diminishing marginal effect again!—and the whole system moves toward a new equilibrium.

Remember that in any economy there are always disturbances and uncertainties at work, so that a precise equilibrium is never reached. Yet changes that move the economy away from equilibrium are automatically met by forces that move the economy back toward equilibrium. This automatic pull toward equilibrium is important, since it gives the market system a self-correcting stability.

Public choice

The final microeconomic principle applies to choices by governments. Such choices are important because the private market system frequently creates *external effects* (often called spillovers or third-party effects) that cause real harm to others, such as when a factory pours out smoke or toxic chemicals as part of its own profit-maximizing choices. If the firm causing the pollution ignores this harm, government action may be justified to correct it.

External *benefits* may also occur, such as from a lighthouse that guides ships, or roads that are open to all, or a police department that protects the citizens. The extreme case of external benefit is called a *public good.* No private firm can profitably provide it because its benefits are widely shared among those who do not pay for it. Such public goods are another economic responsibility of governments.

As governments make the economic choices that treat external effects and provide public goods, those choices can be judged against economic standards of efficiency. The standards are based on *cost-benefit analysis,* which defines the correct level of public regulation and public goods in each case. In precise terms: ***An efficient public action proceeds until its marginal benefits just equal its marginal costs.*** The correct amount of regulation or spending

is chosen in light of the alternative uses of resources. For example, a city's road repair program costs $14 million per year. If it is efficient, the fourteenth million dollars provide at least $1 million in benefits from smoother traffic flows. For another example, the decision whether to build a new public school should weigh the cost of the school against the benefits it will provide.

Macroeconomic principles

Macroeconomics has the same workaday subject matter as microeconomics: firms and households, inputs and outputs, demands and supplies, prices and incomes. But instead of looking at economic life up close, unit by unit or market by market, ***macroeconomics*** looks from afar at the broad outlines of the economy. Microeconomics is more interested in the detail of economic life. But if you have ever studied or visited a great Gothic cathedral, you know that to appreciate the genius of its creators, you cannot just concentrate on the altarpiece, the stained glass, and the sculptures. You must also concentrate on the beauty of its overall conception.

Strictly speaking, macroeconomics has few principles of its own. Its main results come from applying the same principles and methods of analysis used in microeconomics, such as comparative advantage or equilibrium, to the collective behavior of the large interdependent groups that make up an entire economy.

The contrast between the two major branches of economic science can best be brought out by comparing the kinds of questions each branch focuses on. Here are some examples:

Micro: Why have food prices been rising faster than clothing prices?

Macro: Why did prices in general rise more rapidly in the 1970s than they did in the 1960s?

Micro: Why are the wages of blacks so much lower than those of whites?

Macro: Why did the general level of wages (in real purchasing power) fall throughout most of the 1970s?

Micro: Why is the unemployment rate in Michigan so far above the national average?

Macro: Why did the national unemployment rate rise so sharply from 1973 to 1975?

In each of these pairs of questions, the first question applies to a part of the economy, the second to the economy as a whole.

Since macroeconomics deals with the economy as a system, its central concept is *interdependence.* The economy is not simply the individual unit writ large. It is a system of people and groups that interact. Without appreciating these interactions, you cannot understand the ebb and flow of the business cycle or the long-term trends in production and prices. The logical starting point for macroeconomic analysis, then, is with the most graphic description of this interdependence: the circular flow.

The circular flow

Think back to the circular flow whose pattern you saw in Figure 2. It charts the flow of economic activity—goods, services, and money. Households supply inputs—their land, labor, and capital. In return, they get wages for their work and property income from their land and capital. Firms use this labor and property to produce goods and services. Some of these products are destined for other firms. They may use them as intermediate goods, as steel is used as an input in the production of cars. Or they may accumulate these goods as capital—machinery, inventories, and buildings that are necessary to their processes of production. Other goods are destined for con-

sumer markets where they will be bought by households.

As this dual circular flow of goods and money payments occurs, some parts of it must always balance, while other parts balance only if the economy is in equilibrium. *These identities and equalities make up the main substance of macroeconomics as a body of theory.* As you review these principles in the sections that follow, try to keep firmly in mind the central concept of ***interdependence,*** which means that a change in any one part of the circular flow must cause changes elsewhere.

Income and output

To see more clearly what occurs in the dual circular flow of goods and money, consider a simple purchase like a $20 shirt. First, think of the chain of goods and services that went into the production of that simple commodity: fibers, plastic for buttons, labor, machinery, electricity, fuel, rent, profit, and so on. The actual list of inputs for even this simple item would be quite long.

Now think of the money side of the flow, the $20 that you paid for the shirt. Starting with the retailer, it makes its way back down the chain of suppliers. For example, the retailer may pay $5 of this $20 to employees as wages, $10 to the shirt manufacturer, $3 to other suppliers such as the utility companies or landlord, and have $2 left in profit. Now the shirt manufacturer may pay $3 to its employees, $5 to its suppliers, and keep $2 in profit. As you trace this $20 back through the circular flow, the wage and profit figures keep mounting. In fact, if you traced the payments back through all the producers involved in the production of the shirt, the accumulated wages, rent, interest, and profits would mount until they equaled $20, or the entire value of the shirt.

This illustrates one of the most important facts of macroeconomics— ***all expenditures on goods generate an equal amount of income in the form of wages and property income.*** The total value of all goods, then, must equal the total income received from their production.

This equality is an *identity.* It must hold true, regardless of what is happening in the economy. Whether the economy is doing well or poorly, whether it is in equilibrium or not, the value of output will equal the income generated by its production.

Demand and income

From studying the circular flow, you already know where manufactured goods go: to households and firms. But what happens to the income generated by production? It also goes to households and firms. Since wages are paid to employees, they become part of *household income.* So do rent and interest. Profits may be paid to the owners of businesses, so that they become part of household income, or they may be retained by the firm as business saving. Thus, all wages and part of profits become household income. Some of this income is spent by households; some is saved.

Firms invest in plant, equipment, and inventory, and they must somehow raise the money to pay for these goods. Part of these purchases are financed by business saving, but a large part is financed by business borrowing. It should be obvious that the business sector can borrow only what the household sector is willing to save. And somehow, funds must be channeled from the savers to the borrowers. This transfer of funds is the job of the *financial markets.* In exchange for lending their money, savers get title to *future* income from stocks, bonds, savings accounts, and similar assets.

Since saving and investment decisions are made separately, there is no obvious reason why the amounts that households

collectively wish to save must equal the amounts that firms collectively wish to borrow. Because savers and investors are different people, facing different problems, looking at an uncertain future from different perspectives, the amounts people try to save and invest can differ from one another. *When saving and investment plans do not match, desired or planned demand will not equal production.* Suppose that firms want to invest *more* than savers want to save. If planned investment is greater than saving, then planned demand (planned investment plus consumption) will be *greater* than income (saving plus consumption). Remember that income must equal actual output. Thus, if planned demand is greater than income, it must also be greater than actual output.

By the same token, if firms want to invest *less* than savers want to save, then planned demand (planned investment plus consumption) will be *smaller* than income (saving plus investment). Since income must always equal output, then demand must be less than output. Thus, only if the plans of savers and borrowers harmonize will demand and production be equal.

But what happens if demand and output do not match?

Changes in output

Start with the case in which planned investment is less than saving, so that planned demand (investment plus consumption) is smaller than income from production (saving plus consumption). Firms will be unable to sell their output, and producers will accumulate unwanted inventories of finished goods. They will want to cut back on their output, of course, and perhaps on prices too. But think of what happens when production falls to meet demand. Since income and production must be equal, income will fall along with production. With less income, households will consume less, and firms will find it a poor time to invest in expanding their capacity. Demand will fall still lower. What will happen is that the downward movement of production to meet demand will trigger further drops in demand, and therefore production, that are referred to as the downswing of the business cycle. (The popular term is *business cycle*, even though the swings do not occur with clockwork regularity.)

This whole process also works in reverse, of course. If planned investment is greater than saving, planned demand will be greater than output. Firms will find inventories being drawn down and customers being turned away. Excess demand will make them want to increase both output and prices. As production increases, income will rise and households will want to consume more. Firms will find that it is a good time to invest in more capacity. Thus, as output increases to meet demand, the upward movement of production to meet demand will trigger further increases in demand, and therefore production, that are the upswings of the business cycle.

But is this equilibrium level of output and income a "good" one? Does it represent a healthy economy, or one in which production is so low that unemployment is a worry, or so high that shortages make constant price increases a worry? To judge how well the economy is doing, you need to compare the equilibrium level of output with *potential output.*

Potential output

In the microeconomic section of this chapter, you encountered the *production-possibility boundary*, the outer limit to the combination of goods that an economy can produce with its resources. Macroeconomics has its own nearly identical concept: **potential output, *the maximum value of all goods and services that the economy can produce without generating shortages and widespread inflation.*** In other words,

it is the amount of output that the economy can produce without straining its capacity.

When the upswing of the business cycle carries economic activity past potential output, shortages develop, causing wages and prices to rise. As the downswing of the business cycle moves the economy away from potential output, the strain on the economy is alleviated, and wage and price inflation subsides. The cost, however, is unemployment and excess productive capacity. The ups and downs of the cycle sometimes carry the economy hard up against the production-possibility boundary, where there is genuine scarcity, and sometimes to the interior of the curve, where there is waste and unemployment.

A nation in good macroeconomic health is always close to its potential, so that it wastes very little, but is never beyond its potential long enough to generate a chronic pattern of rising prices.

Stabilization policy

The ups and downs of the business cycle, with their alternating episodes of inflation and unemployment, are among the most distressing aspects of economic life. While all economists agree that these episodes should be avoided, there is no single remedy that every economic physician would prescribe. Indeed, how best to stabilize the economy at a level of income close to potential output is one of the major controversies of modern economics.

One group of economists feels that it is the government's responsibility to curb the upswings and downswings of the business cycle. These "liberal" economists believe that the government ought to pursue an active macroeconomic policy. It should stimulate the economy in bad times by spending more and promoting easier credit, and should curb the economy in prosperous times by reducing government spending and tightening credit. These economists feel that such a government ***stabilization policy*** will improve the functioning of the private economy.

On the other side are the more "conservative" economists, who regard the government itself as a major source of economic instability. These economists believe that the government should simply balance its budget at modest levels, rather than trying to stimulate or contract the economy by changing the size of the budget. Similarly, they would urge the government not to influence the level of income by manipulating interest rates. They would like the government simply to ensure that the supply of money and credit expands at some steady rate consistent with overall price stability.

To these economists, nearly all governmental attempts to stabilize the economy are poorly timed, and the lags in the economy's reaction to change are too uncertain. The result is that policies aimed at reversing the downturn will take effect too late and simply overaccelerate the inevitable upturn.

The participants in this debate generally agree on goals: to avoid unemployment and inflation. But the two groups are far apart on how to reach these goals. Like all major scientific controversies, it is a heated dispute, and one, moreover, in which everyone in the economy—yourself included—has a stake, since we are all affected one way or another by the cycle of prosperity and depression.

Macroeconomics, therefore, focuses on issues that are of great importance to all of you. The debates are often sharp and lively. Even in the heat of argument, however, no competent economist should exchange logical analysis for impassioned rhetoric (although logical analysis can certainly be lively!). Like microeconomics, macroeconomic analysis must be constructed carefully, concept by concept. Its major subjects are: ***interdependence; the***

necessary equality of income and output; how the relation between savings and investment affects output; the desirability of coming close to potential output, and finally, the controversies centering on how the right level of economic activity can be achieved.

Summary

This chapter takes a detailed look at three aspects of economics: the U.S. economic system as illustrated by the circular flow, and the two major branches of economics, microeconomics and macroeconomics.

1. The two basic units of the economy are households and firms.
2. Firms and households meet and interact in markets: groupings of buyers and sellers of a good. In a properly functioning market, the price adjusts until the buyers and sellers want to exchange the same amount of the good.
3. The underlying circular pattern of the economic system can be clearly seen by examining the circular flow diagram. This shows the dual flows of goods and money that circulate among the units of the economy.
4. When economists deal with the individual sectors or markets that make up the circular flow, they are working in the branch of economics called *microeconomics.*
5. The *production-possibility boundary* is a diagram that shows the maximum combination of goods that can be produced, given the economy's resources and the known technology.
6. *Macroeconomics* is the branch of economics that considers the economic system as a whole, rather than the individual units or markets of the economy. Interdependence, meaning that a change in any one part of the circular flow must cause changes elsewhere, is the key to macroeconomic analysis.

Key concepts

Enterprise
Inputs; factors of production
    labor
    capital
    land
Opportunity cost
Marginal analysis
Diminishing marginal effect
Efficiency
Production-possibility boundary
Household
Maximizing
Market
Circular flow diagram
Macroeconomics
Microeconomics
Specialization by comparative advantage
Equilibrium
Potential output
Stabilization policy

Questions for review

1. You decide to spend more of your money on clothes and less on food. Chrysler introduces a new line of compact "K" cars, and cuts back on production of larger cars. Can you see any common motive that might explain both actions? Discuss.
2. Economists talk about "constraints" that limit the choices of consumers and firms. List the constraints that you face as you make your own consumption decisions.

3. Why is the circular flow sometimes referred to as a *dual* flow?
4. Pick any three decisions that you had to make today. What was the opportunity cost to you of these decisions?
5. Explain how diminishing marginal effect limits your own consumption of some particular good.
6. Consider the following situation: The UAW calls a strike against GM. Explain the chain of reactions that will take place throughout the economy, showing the interdependence of the units of the economy.

3

Methods and Measurements

As you read and study this chapter, you will learn:

- what economic models are
- how graphs can show economic patterns
- how to guard against deceptive graphs
- the main accounting concepts of flows and stocks
- how economists treat specific issues

Occasionally, if you are a regular reader of the Sunday paper, you will run across an article about someone who has made a 6-foot model of the Eiffel Tower out of toothpicks. The builder is usually in prison, with a lot of time to serve. Another favorite hobby for those with time to kill is collecting the "world's biggest" ball of twine—although you usually can't work on this in prison. Still other people seem to enjoy calculating π to an immense number of decimal places. This preoccupation with activities that have no purpose beyond themselves is a uniquely human trait. Many animals are playful, but only men and women seem able to waste their time in deadly earnest.

The academic equivalent of the 6-foot Eiffel Tower is found in the higher reaches of pure mathematics. To work comfortably at this intellectual altitude, a person must be in love with logic for its own sake.

Economics differs from pure mathematics. Though economic concepts have a certain logic of their own, they ultimately matter only if they can clarify the complexities of the real world. You probably had practical objectives in mind when you decided to study economics. Practicality, in fact, is the hallmark of economists, as well as of economics. Nearly all of the great economists have had their feet planted firmly in the world of affairs—whether business, politics, or social revolution. Economists are also incurably empirical—they like to learn from facts. They have always tried to fit and apply their concepts to the facts of economic life.

Empiricism requires measurement. Yet, precise measurements are often difficult to make, even in the dollars-and-cents world of economic life. Every crucial condition—such as unemployment, price trends, costs, total production of goods, profits, pollution, or the benefits of public programs—can usually be measured in several ways, each of which often gives different results. Moreover, cause and effect are often muddled, even after decades of research. Do two variables fluctuate together because one causes the other to change? If so, which is the cause and which is the effect? Or, do both change in response to a third variable?

To complicate things further, major economic issues are sensitive matters that involve large stakes. They attract partisans of vested interests, who often make their points with biased data and deceptive charts. You have already been bombarded for years with debatable "facts" about the economy from television, magazines, newspapers, and other sources. The economist's task is to sift the reliable facts from this outpouring of data, and to discard the trivial and the distorted. Printed numbers, like printed words, are fallible. Often, the more exact a published "fact" seems to be, the more distorted it really is.

This chapter will show you how basic economic ideas can be applied in practical forms. If you are already comfortable with graphs and linear equations, you can pass quickly over some parts of this chapter. But if graphs and equations seem daunting, then this chapter can help to overcome your aversion. Read it slowly and carefully. You will discover that diagrams and measurements can make economics much easier to understand. Take your time with the material. When you finish, you will be acquainted with every mathematical concept you will need in order to understand the whole text.

The first section deals with graphs and their uses. First, we present simple linear or straight-line equations and their graphs. Then, we use this information to show how a simple economic model can be constructed. Next, we discuss other types of graphs often used in economic analysis—time series and distributions. The second section shows how to interpret numbers and graphs to avoid problems of bias and deception. Finally, we present the distinction between stocks and flows.

Diagrams and their uses

To relate economic theories to the data and facts of the real world, economists often rely on diagrams. If diagrams intimidate you, try to think of them simply as cartoons, useful for illustrating ideas and facts in a simple way. They are used widely in economics to convey the sense of words and numbers. Although good economic analysis can be expressed clearly in plain words, diagrams (also called graphs, figures, or charts) can often portray facts and relationships with greater clarity.

Suppose, for example, that you are told that people's total savings (1) increase as their after-tax income increases, (2) decrease as income decreases, (3) are zero at

some level of income, and (4) can even become negative. It would take you some time to sort out and assimilate all that information. Even those of you with only a slight acquaintance with graphs would probably find this information much easier to digest if it were accompanied by a diagram. A graph makes the information about the income-saving relation both easier to understand and easier to remember.

Don't fall into the trap of skipping over diagrams and hoping that they don't matter. Study them carefully and become skilled at drawing them. Practice drawing different versions of them over and over. Use diagrams to clarify your own thoughts when you are studying or taking exams. You will soon see how much they add to your understanding rather than to your confusion.

There are three main types of diagrams that can be used to convey many different kinds of economic concepts: (1) graphs that present a simple *economic model* showing the relationship between two variables or concepts—between price and quantity, for example, or consumption and income; (2) graphs called *time series* that show how an economic variable (such as prices, unemployment, or national output) has changed over time; (3) finally, graphs called ***distributions*** that show how an economic variable such as income or wealth is spread throughout the population. Does a small percent of the population have most of this country's wealth, or is it fairly evenly distributed among the population? A distribution can give the answer.

First, let us discuss the type of diagram that shows the relationship between two variables. The emphasis will be on those relationships that can be shown as a straight-line or *linear relation*. First, you will see how to interpret and graph a linear relation. Then, using this knowledge, you will learn how to construct a simple economic model. Those of you who have not worked much with linear equations and graphs will probably be surprised at how clear and simple it is to use them. After dealing with the equations and graphs of simple economic models, we will then take up the two other major types of graphs used in economics: time series and distributions.

Linear equations and their graphs

Many economic theories merely relate two quantities. Often, the relationship can easily be pictured or approximated by a straight-line graph. In Figure 1, for example, the straight line summarizes the possible relationship between people's consumption and their after-tax or disposable income. By studying the graph, one can see how consumption changes when disposable income changes. As shown, when disposable income changes by $100, consumption changes by $80. This graph of the income-consumption relation can be described or represented by a ***linear equation.*** A linear equation has the general form of: $y = a + bx$, where a and b are constant numbers (or coefficients). The ***independent variable***—the cause—is represented by x. This is the variable that is thought to cause changes in the ***dependent variable,*** represented by y. In the consumption-income relation, changes in income are believed—on the basis of logic and the observation of real behavior—to cause changes in consumption. In the linear equation representing this relation, x would represent income, while y would represent consumption.

When the straight line representing the equation is graphed, the convention is to put the independent variable or x on the horizontal axis, while the dependent variable or y appears on the vertical axis. (In economics, there is one major exception to this: the graphing of demand and supply curves showing the price and the quantity

Economics Requires Measurement as Well as Theory

Every year since 1969 a Nobel Prize in Economic Science has been awarded to one or two economists, who are shown below. Many of these preeminent economists have focused on the measurement of economic conditions, developing the tools of *econometrics*. All of the Nobel winners have dealt extensively with economic facts.

Winners of the Nobel Prize for Economics

1969
Ragnar Frisch
Norway (L)
Jan Tinbergen
Netherlands (R)



For pioneering work on mathematical economics and econometrics, developing models that could test economic concepts with real-world facts.

1970
Paul A. Samuelson
United States



For varied contributions which have raised the level of scientific analysis in economic theory. He also writes widely on applied economic issues.

1971
Simon S. Kuznets
United States



For pioneering work in developing the measurement of economic activity and analyzing the causes of economic growth.

1972
Kenneth J. Arrow
United States (L)
John R. Hicks
Britain (R)



For advancing the theory of general equilibrium and demonstrating special problems that can arise in complex economic systems.

1973
Wassily Leontief
United States



For the development of the input-output technique to portray and analyze interdependencies within the economy.

1974
Friedrich A. von Hayek
Britain (L)
Gunnar Myrdal
Sweden (R)



For their contrasting approaches to economic problems. Myrdal developed criteria for stabilizing and promoting growth, while Hayek urged against government intervention.

1975
Leonid V. Kantrovich
Soviet Union (L)
Tjalling C. Koopmans
United States (R)



For fundamental analysis of the optimal allocation of resources in planned economies (Kantrovich) or market systems (Koopmans).

1976
Milton Friedman
United States



For research stressing monetary conditions as the crucial economic factor in business cycles, and for his influence upon monetary policies.

1977
James E. Meade
Britain (L)
Bertil Ohlin
Sweden (R)

For path-breaking contributions to the theory of international trade and international movements of capital.




1978
Herbert A. Simon
United States

For exploration of the goals, structures, and decision-making processes within large organizations.



1979
Arthur W. Lewis
Britain (L)
Theodore W. Schultz
United States (R)

For research on world poverty, economic development, and the economic yields to human capital.




1980
Lawrence R. Klein
United States

For developing econometric models to analyze and predict fluctuations in economic activity.



1981
James Tobin
United States

For research on the relations between financial and real markets, as they affect aggregate economic activity.



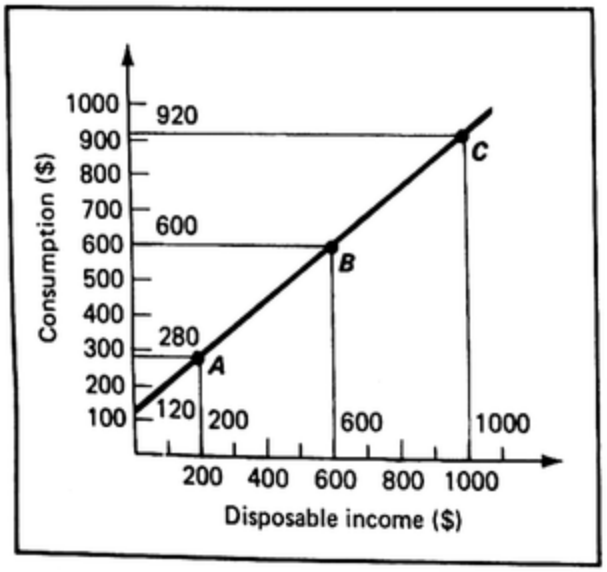


Figure 1 A diagram of the consumption–disposable income relation

The formula is:

$C = 120 + 0.8\ YD$
where C = consumption
and YD = disposable (or after-tax) income.

Point	Value of Disposable Income	Value of Consumption
A	200	280
B	600	600
C	1000	920

of a good that is purchased or sold. Alfred Marshall put price, the independent or x variable, on the vertical axis, and this has become the unshakable tradition. In all other diagrams, economists put the independent variable on the horizontal axis where it belongs.)

The equation of a straight line must contain enough information to draw the line correctly. This information is conveyed by the constants a and b. To see why, look again at the equation

$$y = a + bx.$$

If $x = 0$, then $y = a + b\,(0)$, or simply $y = a$. If you look at Figure 1, you can see that when $x = 0$, the line cuts the y axis at a value of 120. This point is called the ***intercept.*** So for this line, the value of a (or the intercept) is 120.

Now for the slope of the line, look again at the general form of the linear equation. If x increases by 1, by how much will y increase? Try some numbers. Since a is constant, you can see that if x increases by 1, y will increase by an amount equal to the value of b, which is the coefficient of x. The value of the coefficient b is 0.8. So b represents the $\frac{\text{change in } y}{\text{change in } x}$ or $\frac{\Delta y}{\Delta x}$. (Read it as: delta y divided by delta x.) By definition, $\frac{\Delta y}{\Delta x}$ is the slant or ***slope*** of the straight line. So, knowing the value of a gives us the height of the line for $x = 0$. Knowing the value of b gives us the slope or tilt of the line.

That is all the information you need to draw any straight line. For our income-consumption diagram, the linear equation describing the graph is:

$$y = 120 + 0.8\,x.$$

Because b, the slope of the line, is constant, the relationship has a straight-line or linear form. If b were variable, the line would have a genuinely curved shape. The shape of the line—straight or curved—depends, of course, on what the data tell us about the actual relation between the variables.

The values for a and b may be large or small, positive or negative. A positive value for b means that x and y move in the same direction: More x causes y to rise. This results in a line with an upward or *positive* slope. The graph of the income-consumption relation shown in Figure 1 is a good example of such a *direct* relation.

A negative value for b means that x and y move in opposite directions. As x increases, y will decrease. This results in a line with a downward or *negative* slope. An example is a demand curve, relating prices to quantities bought. As a good's price increases, consumers usually want to purchase less of it. The result will be a line with a *negative* slope, similar to Panel II of Figure 2.

In fact, Figure 2 contains four different variations of linear equations. For each equation, a few x and y values are given. Make sure that you understand how to calculate the y values, given the x values. Make sure also that you could have sketched in each line, given the actual values of a and b for each equation.

Of course, the relations between variables may be curved, not straight. Many are. Figure 3 gives two examples. Those two curves will often be found later in this book. Panel I shows a typical total revenue curve, while Panel II's curve is an average cost curve.

By this time, you should be starting to feel comfortable with the meaning of linear equations and with the graphs that portray them. Now it is time to move on to the next step, which is to use this skill to construct a simple economic model.

Economic models

Economic analysis is mainly a set of theories about cause and effect: "Condition A makes result B occur." These ideas are often given exact form as ***economic models. An economic model is simply a precise formal statement of one or more economic relationships.*** The idea begins as words, of course, but it is often put in equation form. The model can be as small and as simple

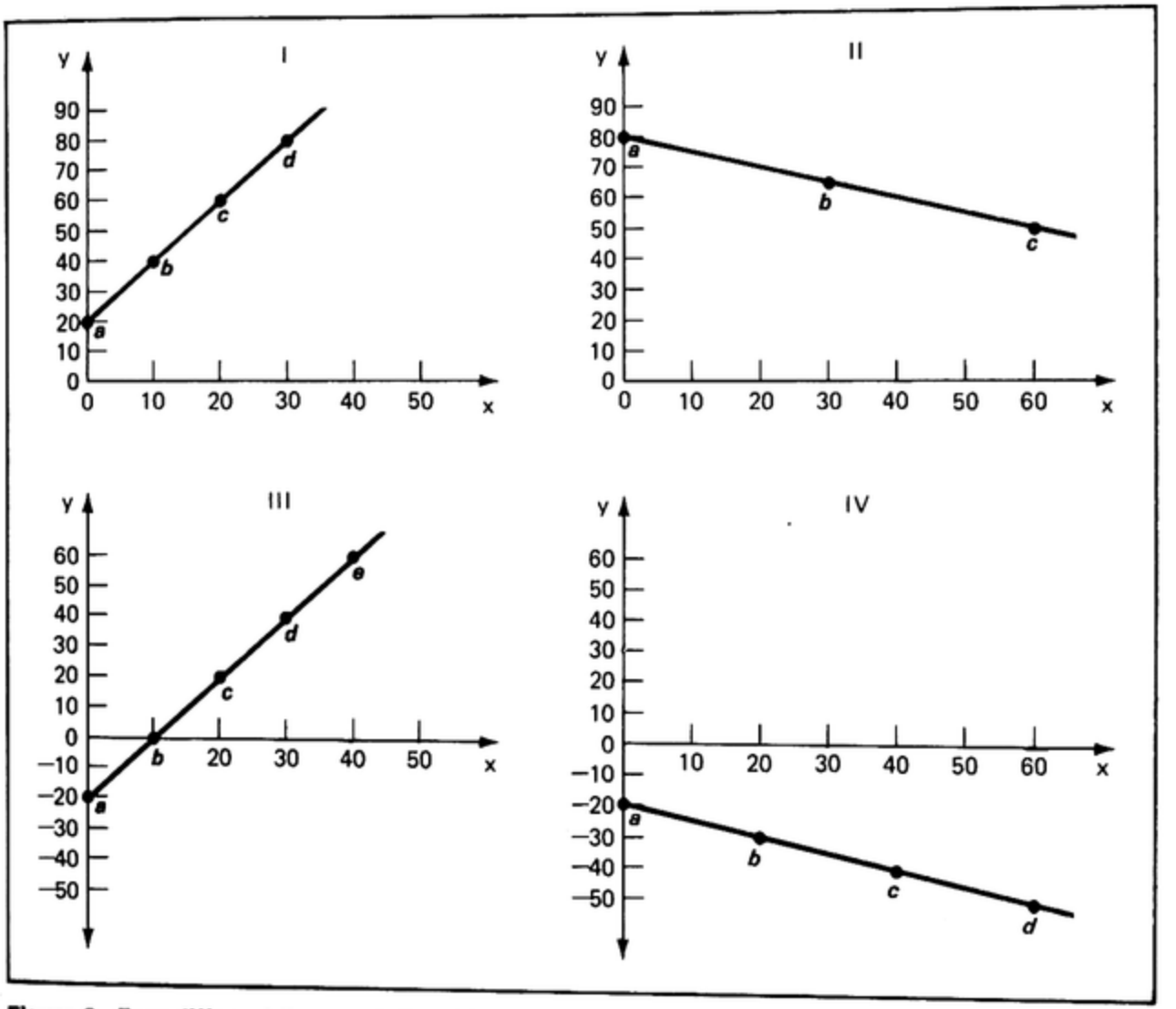


Figure 2 Four different linear relationships

I. This relationship is a line with the formula:

$$y = a + bx$$
$$= 20 + 2x.$$

Point	Value of *y*	Value of *x*
a	20	0
b	40	10
c	60	20
d	80	30

The line has a *positive* intercept and a *positive* slope.

II. For this line, the formula is:

$$y = a - bx$$
$$= 80 - .5x.$$

Point	Value of *y*	Value of *x*
a	80	0
b	65	30
c	50	60

The line has a *positive* intercept and a *negative* slope.

III. This line's formula is:

$$y = -a + bx$$
$$= -20 + 2x.$$

Point	Value of *y*	Value of *x*
a	−20	0
b	0	10
c	20	20
d	40	30

The line has a *negative* intercept and a *positive* slope.

IV. Here the formula for the line is:

$$y = -a - bx$$
$$= -20 - .5x.$$

Point	Value of *y*	Value of *x*
a	−20	0
b	−30	20
c	−40	40
d	−50	60

The line has a *negative* intercept and a *negative* slope.

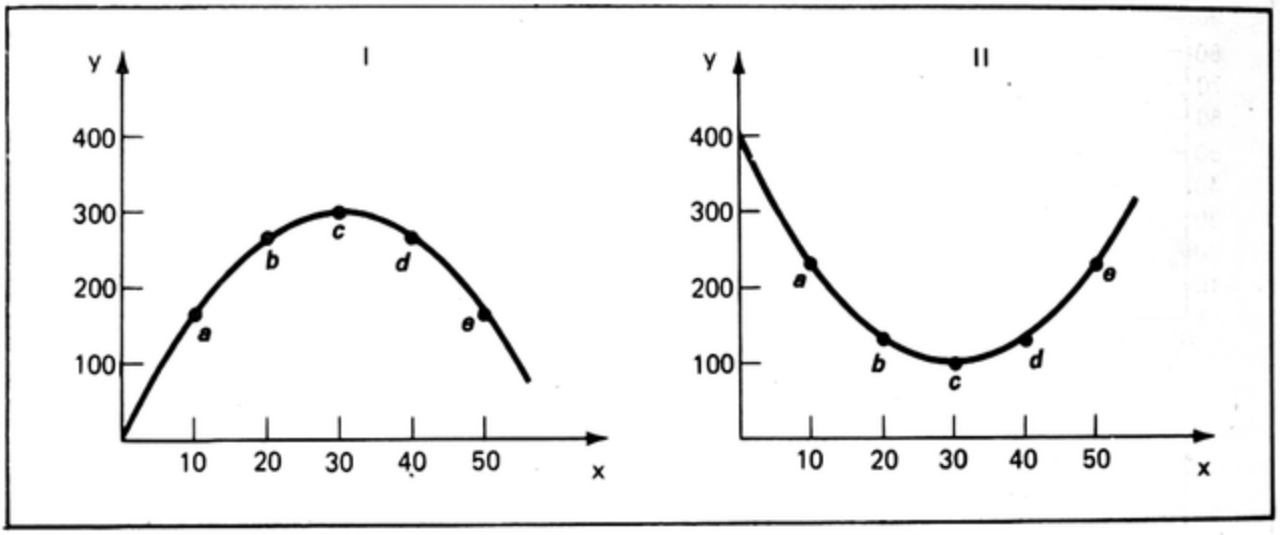


Figure 3 Two illustrative curves

In Panel I, the relationship is:

$$y = a + bx - cx^2$$
$$y = 0 + 20x - .33x^2.$$

The five points shown are:

Points	Value of x	Value of y
a	10	167
b	20	267
c	30	300
d	40	267
e	50	167

In Panel II, the formula is similar, but the signs of coefficients b and c have been reversed:

$$y = a - bx + cx^2$$
$$y = 400 - 20x + .33x^2.$$

The five points shown are:

Points	Value of x	Value of y
a	10	233
b	20	133
c	30	100
d	40	133
e	50	233

as a single short equation relating two variables. For example, a demand curve is a model relating the price of a good to the quantity of it that people buy. Models can also be big and complex, with many long formulas containing dozens of variables. Macroeconomic models are often that large. Yet even the biggest economic model is a simplified version of the real world—a set of ideas about how economic conditions relate to one another.

Suppose that you decide—as seems logical—that a change in disposable income will cause a change in consumption. Disposable income (which we will label *YD*) is then considered the *independent variable*, which causes or explains the change in consumption. Consumption (or *C*) is the *dependent variable*, since it depends on or is determined by income. Trying to think logically about this income-consumption relation, you could well reach two other tentative views. First, you could decide that the income-consumption relation can best be expressed as a simple *linear* relation shown by a straight line. You could then present the idea—or theory or model—as a linear equation of the $y = a + bx$ type. Second, you could conclude that the logical relation between income and consumption is a *direct* one, with consumption increasing as income increases and falling as income falls. That is surely logical. This means that the straight line that shows the relation between income and consumption will have an upward slope.

Now, starting with the general form of a linear equation:

$$y = a + bx$$

you can substitute actual variables and as-

sumptions into this general form and arrive at a linear equation specifically designed for your model. Consumption (or C) replaces y as the dependent variable. Disposable income or YD replaces the more general x as our independent variable. The slope of the line (or b) should have a positive sign because of the direct relation between consumption and income. Finally, you would expect a, the intercept, to have a positive sign because consumption can never be negative. So, your linear model can now be expressed as:

$$C = a + b\,YD.$$

Notice how the logic of the model—the cause-effect relation—is highlighted by having the cause (or independent variable) always on the right-hand side of the equation and the effect (or dependent variable) on the left. The graph of this line would take the general form shown in Figure 1. The independent variable is on the horizontal axis, and the dependent variable is on the vertical axis.

At this point, you should pause and ask if the model fulfills the first important criterion by which the soundness of any model should be judged: Is the model *($C = a + b\,YD$) logically valid?* It seems to be, since the assumptions of the model make sense: From what we know, people do buy more when their income rises. Moreover, the logic of the model does not appear to have inner conflicts.

You can now proceed to the second step: How well does the model fit the facts of the real world? Now you move from theorizing to data collecting and processing: from points of *logic* to matters of *degree.*

When you plot actual observations of consumption-income combinations on your graph, the result may be a "scatter" of points, as in Figure 4. They appear to reflect a pattern. A line may now be fitted to the dots, to suggest what the relation between consumption and income is. This is a tiny example of what *econometrics* does: testing and fitting relationships with real-world facts. You can draw free-hand a line that fits the data points, or you can use computers to fit the line precisely, by statistical methods. In either case, following good statistical techniques, you first collect all the data necessary to derive or "fit" the linear relationship. This task of data collection is often much more difficult than it sounds, for you need to collect data not just on quantity and income, but also on all the other "noise" variables, such as population, that you would want to hold constant.

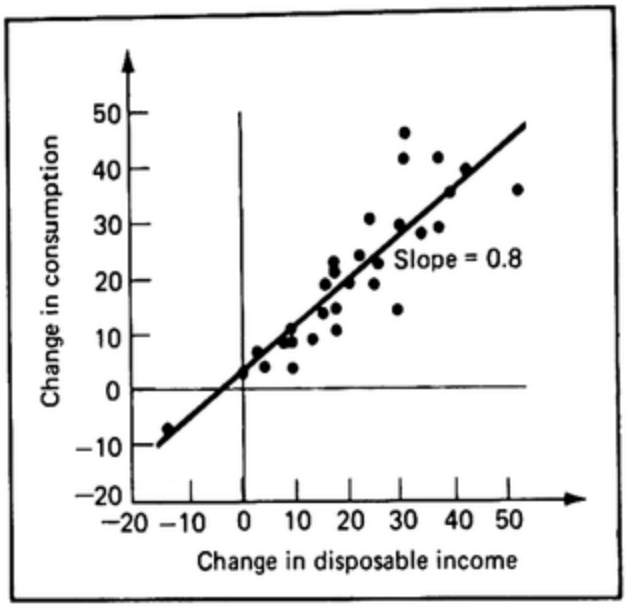


Figure 4 Annual changes in consumption and disposable income 1950–1979 (in billions of 1972 dollars)

This figure relates changes in total national consumption to changes in total disposable income. Each dot represents one year's data. The dots correspond to actual data for the U.S. economy during the 1950–1979 period. A straight line with a slope of 0.8. approximates the data fairly well. By fitting curves or straight lines to historical data, economists are able to measure how some economic variables change in response to other variables.

At this stage, you might decide that although the model made logical sense, it is simply not borne out by the facts. In that case, you would want to reconsider either the model or the soundness of your data

and statistical techniques. Perhaps income was too weak an influence on consumption to have any clear influence on quantity. Perhaps the data were inaccurate. Or perhaps the relation between income and consumption could be better expressed by a curve instead of a straight line. In any case, the model must be modified or the data refined.

If the model does appear to fit and to explain the facts of the real world, it has passed the second important criterion by which its soundness can be judged: *It is testable and consistent with the facts.*

Once you have fitted your model with actual data, you have actual values for both *a* and *b*. For example, the model may now have these values:

$$C = 120 + .8\ YD.$$

Not only do you have a precise definition of the relation between consumption and income, but you can also use the model to predict future changes in consumption as income changes, other things being equal. For example, given the above equation, a $1,000 change in families' disposable income will, on average, lead to an $800 change in their consumption

$$(b \text{ or } .8 = \Delta\ C/\Delta\ YD).$$

There is, however, one more criterion of a sound model to consider: *completeness.* Real markets have many crosscurrents: Five, ten, or even more influences may be at work on one variable. Total consumption clearly depends on the size of the population and on incentives to save, as well as on income levels. Your model need not include every such possible influence. A simple model like $C = 120 + .8\ YD$ is complete enough, if your major interest is finding out how responsive consumption is to changes in income. But if your aim is to predict consumption precisely, a simple two-variable model will not do. You will need a much larger and more complex model, perhaps with ten or fifteen independent variables, allowing for every likely influence on consumption.

An important factor in constructing useful models is to develop a sense for the main economic forces. Always guard against theories that are either too simple or needlessly overcomplicated, given the purpose for which they are being constructed. Beginning students in economics sometimes scoff at models for not accurately picturing all of the world's complexities. Yet, like Newton's three laws of motion, such simple models may be both valid and powerful. They do not explain all details, but they contain the essence or core, the starting point for more detailed study.

So far, the discussion has concentrated on graphs portraying the relation between two variables. It is time to examine the two other types of graphs mentioned earlier: time series and distributions.

Time series

A ***time series*** shows how an economic variable has moved or behaved over time. A suitably long time series can be remarkably helpful in clarifying the basic pattern of growth or decline. Short time series, however, are often suspect, since they show only the latest shifts, which are difficult to interpret out of their historical context.

Usually, the vertical axis of a time series is an ordinary number scale, as in Figure 5, graph A. Equal increases are represented as equal distances. But if the growth process is believed to be a steady percentage rise that cumulates or steadily increases over time, such as the growth in national output, you might use a ratio scale on the vertical axis, as in Figure 5, graph B. (It is also called a logarithmic scale.) With a logarithmic scale, equal ratios are represented as equal distances. With such a scale, a constant percentage

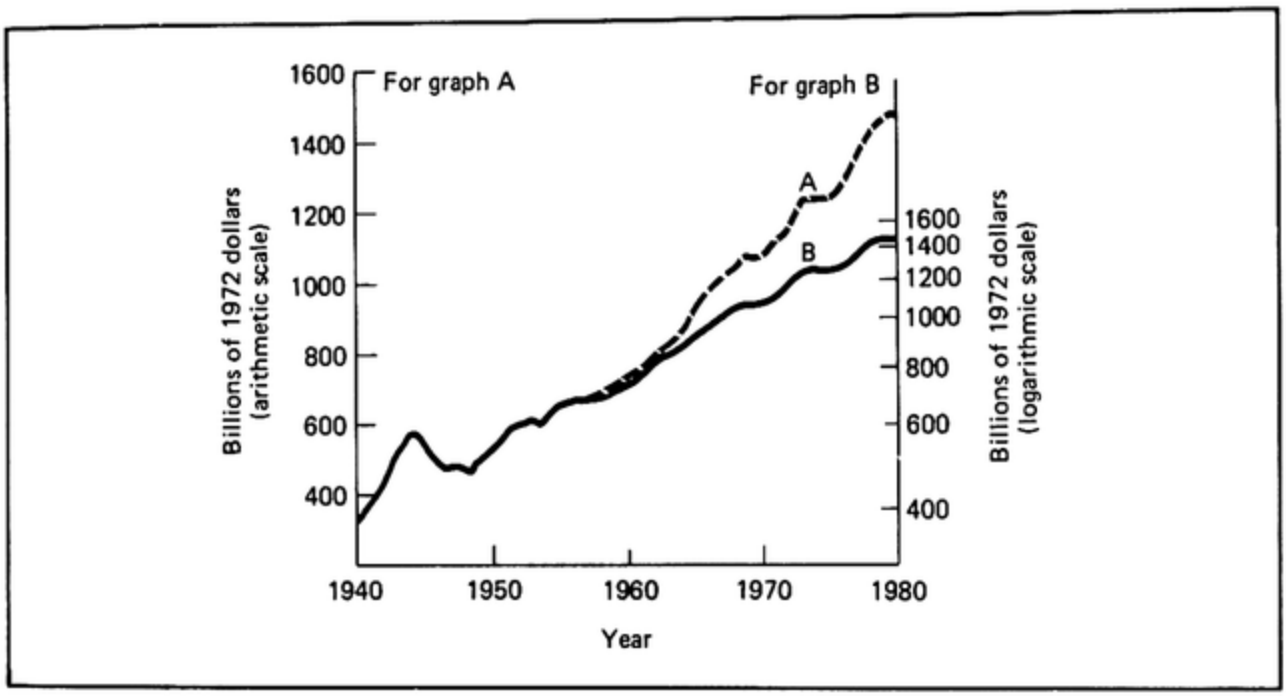


Figure 5 Two ways to show a time series

Graph A and Graph B give the same information. They show the path of U.S. GNP (in constant dollars) for four decades. A is to be read relative to the left-hand scale, which is an arithmetic or absolute scale. Equal dollar amounts appear as equal distances. B is to be read relative to the right-hand scale, which is a logarithmic or ratio scale. Equal percentage amounts appear as equal distances. The graphs have been placed so that they virtually coincide in the early decades. But as GNP gets larger in later decades, equal percentage increases show up as ever-increasing dollar increases. B conveys the correct impression of a fairly steady rate of growth over time. A conveys the incorrect impression of an ever-accelerating rate of growth.

rise results in a straight line rather than an accelerating up-slope. The ratio scale is thus a good way to keep a growth process in perspective. Any change in the proportional rate of growth shows up immediately as a change in the line's slope.

The decision to use a number or a logarithmic scale depends on whether the nature of the change itself is cumulative (that is to say, continually increasing). Since unemployment rates rise and fall with no necessary trend, a number scale would be the best choice for them. But for growth that increases national output cumulatively over time, a logarithmic scale would best convey the information. When interpreting a time series graph, try to imagine how the alternative version would look. This will help you to avoid deceptive graphs.

Distributions

Distributions of numbers are often important in economic analysis. They can be presented in a bell-shaped curve, called a ***normal curve.*** Such a distribution shows values clustering symetrically around the average value and then trailing off at the upper and lower ends of the distribution. "Grading on a curve" means following such a bell-shaped distribution, as Figure 6, panel I shows. There is a small percentage of the extremely high and low grades of A+ and E. Most grades cluster around the average grade of C.

Most distributions of economic variables such as income and wealth do not follow this bell-shaped distribution. They usually lean toward or are skewed toward high or low values. The more skewed they are, the more unequally the variables they

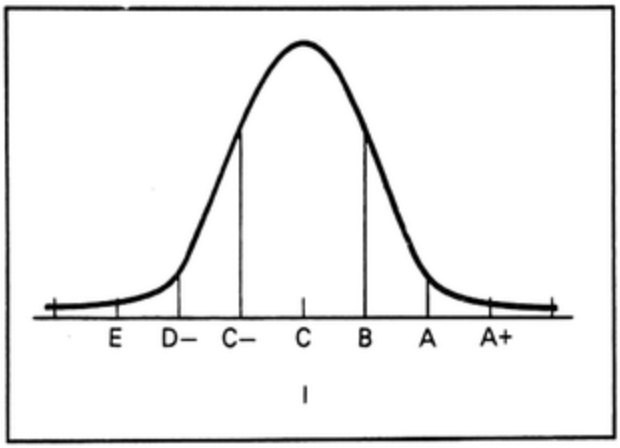


Figure 6 I A normal distribution of grades

Here, letter grades of A+ through E have been assigned to reflect a *normal curve*. Grades cluster around the average value or grade of C. Very few students receive grades of A+ or E, which lie at the upper and lower extremes of the grade range. In this particular class, the distribution is:

Grade	% of Students
A+	2.5
A through B+	13.5
B through C–	68.0
D+ through D–	13.5
E	2.5

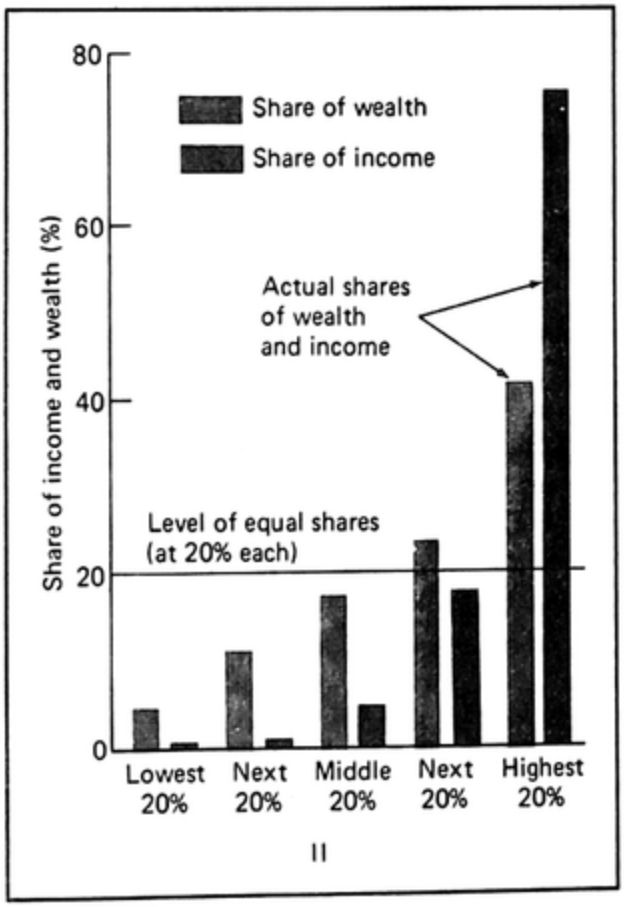


Figure 6 II The distributions of income and wealth in the United States

The population has been ranked according to income and wealth and then divided into five groups. The poorest households are to the left, while the richest households are at the right end of the scale. The distribution shows that the poorest one-fifth of households received 5 percent of all income, while the richest one-fifth of households received 42 percent of all income.

The wealth distribution was even more unequal. The poorest one-fifth of the population had virtually no wealth, while the richest one-fifth of households held 76 percent.

Source: *Statistical Abstract of the United States.*

present are distributed (although any distribution, even if it is symmetrical, involves *some* inequality).

For example, consider the distribution of income and wealth illustrated in Figure 6-II. The distribution is skewed toward the rich. In 1980, the richest 20 percent of the population held 76 percent of the wealth in the United States, while the poorest 20 percent had virtually no wealth. The richest 20 percent of the population received 42 percent of the income, while the poorest 20 percent received only 5 percent. You can see how the distribution shows the inequality of wealth and income in the United States more quickly and clearly than would be possible in words.

The interpretation of numbers

Problems of bias and deception

By now you realize that economists form their concepts in words and test them with numbers. Economists measure many things, such as prices, trends in output, and relationships between variables such as income and consumption.

Yet numbers by themselves do not convey meaning. They are interpreted and presented by people, often in ways that can twist or bias their message. Graphs are especially liable to deception. One must treat every graph with caution. Even an

TEXT BOOK LIBRARY
Acc No 15355

honest graph can deceive, and many graphs are *designed* to deceive.

Graphs can be biased on two levels. *First*, they may contain data that are poor or off the point. An economist usually has several kinds of measures of an economic concept to choose among. Each may give a different impression.

The *second* level of possible bias is the form of the graph itself. Several common deceptions involve shortening the vertical axis, stretching or shrinking the axes, putting a low ceiling on the graph, extrapolating carelessly into the future, and using deceptive forms of bar charts. Though these tricks are well known, they still recur daily in newspapers and magazines. Often the deception is not intentional: The artist merely tried to fit the graph into a limited space or sought to dramatize the point. Indeed, every layout stresses *something*. Remember: There is no perfectly neutral basis for a graph.

Stocks and flows

To interpret numbers carefully, you often have to know if the data refer to a flow or a stock. ***Flows*** refer to processes or values occurring *during a period of time*. For the flow to be understandable, its precise time interval must be stated. For example, knowing that a person earns $5,000, or that a store sold 10,000 pairs of shoes, is not enough. Is the person earning $5,000 a month or $5,000 a year? Did the store sell 10,000 pairs of shoes in a week, a month, or a year? Knowledge of the time element is essential if the volume of a flow is to make sense.

Stocks, in contrast, are concerned with the value *at a point in time*. Thus, a reservoir is a stock of water, while a river leading out of it is a flow of water. A factory is part of the stock of a nation's capacity to produce. Inside that factory, the production activity is a flow.

Table 1 shows the main stocks and flows both of the three domestic sectors of the economy—households, firms, and governments—and of the economy as a whole. ***Note that stocks and flows can be measured in either physical or monetary terms. When you are studying an economic concept, always try to be clear on whether you are studying a flow, a stock, or a relationship between the two.***

Households The members of each household engage in such flow activities as work and consumption. Their actions are measured in *physical terms* such as hours worked per week and numbers of loaves of bread eaten per month. There are corresponding *money* flows, of dollars of income from the work, of dollars spent for consumption goods, and of dollars saved when spending is less than income.

Households have physical stocks of possessions, primarily their houses, land, automobiles, appliances, furniture, and so on. Their monetary stock data have two sides: One side is the ownership value of those physical assets, plus whatever money assets the household has (cash, bonds, stock certificates, savings deposits, etc.). The other side is liabilities, such as debts, loans (house mortgages), and bills owed. The household's net worth is its assets minus its liabilities. Net worth can be negative or positive.

Firms The flows within firms are the volumes of inputs coming in (labor hours, raw materials, and services), the levels of production activity, and the amounts of outputs going out (such as thousands of bushels of wheat, or thousands of tons of copper). These flows are measured in money values, by multiplying their physical volumes times the price of these items. For example, if Ford sold 1,925,000 cars in 1981 at an average price of $7,500, then the sales revenue was $14,437,500,000 per

Table 1 ***Flows and stocks for the three domestic sectors and for the entire economy***

	Flows		Stocks		
	Physical	Monetary	Physical	Monetary	
				Assets	*Liabilities*
Households	Time and effort in work	*Income*	Land, houses, cars, appliances, etc.	Ownership of: *Tangible assets:* land, house, etc.	*Loans* mortgages *Bills owed*
	Time and amounts of consumption activity	*Spending*		*Monetary assets* cash, bonds, stocks, etc.	
Firms	Production Inputs used	*Saving* *Sales revenue* *Current costs* Profit − taxes *Profit after taxes* a. Dividends b. Reinvested in the firm	Land, buildings, equipment, inventories, etc.	*Total Assets* *Fixed assets* land, plant, machinery, etc. *Inventories* *Monetary assets*	*Total Liabilities* *Debts* Long-term bonds Short-term loans *Equity* of stockholders
Governments (national, state, local)	Employees Amounts of activities and services performed	*Investment* *Tax revenues* Income tax Sales tax Property tax Profits tax *Expenditures* Purchases of real goods and services Transfer payments Additions to assets	Land, buildings, equipment, inventories, weapons, roads, and harbors	*Total Assets* *Long-term physical assets* (land, buildings, etc.) *Inventories* (stockpiles of metals, grains, etc.)	*Total Liabilities* *Debt* U.S. government bonds (the national debt) state bonds local bonds *Money* (paper and metal)
Entire Economy (local or national)	Production (items and totals) Employment levels Unemployment levels	*Gross National Product* (GNP) *Sources*, by sectors *Uses* Consumption Government use of goods and services Investment (= to saving)	Land, buildings, equipment, roads, natural resources, inventories, etc.	*Total Assets* *Fixed assets* (land, buildings, etc.) *Inventories* (gold, other metals, other goods) National currencies *Total Assets*	*Total Liabilities* *Ownership claims* *Money supply*

year. If costs included 110,000 worker-years at an average cost per worker of $22,000, then labor costs were $2,420,000,000 per year. Its profit was another flow: total revenues minus total costs.

Firms' physical stocks include the obvious land, buildings, and machinery, plus the inventories of inputs ready for use and of outputs that have not yet been shipped out. These physical items all have their monetary values, such as the Bell System's $100 billion of assets in its present telephone system. Firms also have money assets, such as cash and securities. And firms have debts, in the form of their loans and bonds. The common stocks owned by shareholders are also liabilities.

Governments Still other versions of these flow and stock categories exist for governments. Their flow activities include the employees they hire, the levels of their operations, and the services they provide. The monetary flows have two sides, just as for households and firms. Tax revenues provide an inflow of funds from a variety of taxes on income, sales, property values, profits, and other things. Governments spend to pay for employees, services, and other purposes. As for stocks, governments use large volumes of physical assets, such as land, buildings, stockpiles, and equipment (a police station, a submarine, a road). Governments also issue bonds and money as liabilities.

The entire economy All of these magnitudes add up to form flow and stock values for the whole economy. The physical flows include production, consumption, growth rates, employment levels, and others. These flows are valued in money terms to give the various measures of gross national product, national income, consumption, investment, saving, and many other components.

The economy's physical stocks include the familiar land, buildings, equipment, roads, natural resources, inventories, and other types. They are all valued as assets in money terms, and there are various ownership claims on them.

From the vast economy-wide totals down to minute levels for individual units, the values of flows and stocks all fit into the same basic logic.

Summary

This chapter has focused on the economist's basic tools of measurement and modeling. The major points of the chapter are as follows:

1. Diagrams summarize ideas and facts. They often convey information more clearly than words.
2. Linear equations and their graphs are often used in economics to summarize the relationship between two variables. A linear equation has the general form: $y = a + bx$, where a and b are constant numbers. The constant a represents the *intercept* of the line, the point at which it touches the vertical axis. The constant b represents the degree of slant or *slope* of the line. The letter x refers to the independent variable, the variable that is thought to cause the change in the dependent or y variable.
3. Once the model is constructed, its usefulness should be judged by three important criteria:
   a. It should be *logically valid.* The assumptions should make sense.
   b. It should *fit the facts* of the real world. In other words, it should be testable and consistent with the facts.
   c. It should be sufficiently *complete,* but not needlessly overcompli-

cated, given the purpose for which the model was constructed.

4. A *time series* is a graph that shows how an economic variable has behaved over time. A time series may be presented on a number scale or a logarithmic scale.
5. *Distributions* are graphs that show how such economic variables as income or wealth are spread among the population. The usual distribution shows the majority of cases clustering around an average value, with fewer cases in the upper and lower ranges.
6. To interpret information correctly, it is important to distinguish between stocks and flows. *Flows* refer to processes or values occurring over a period of time. A *stock* is a value at a point in time.

Key concepts

Linear equation:
- dependent variable
- independent variable
- intercept
- slope

Economic model

Time series

Distribution

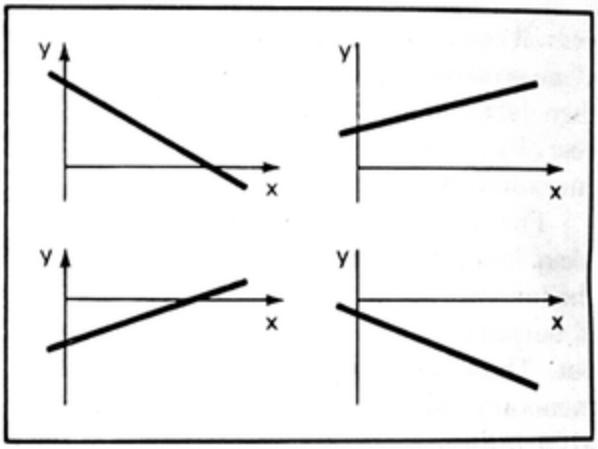
Normal curve

Flow

Stock

Questions for review

1. Suppose that a linear equation has the form of: $y = -10 + .5x$. Using the equation, calculate some values for x and the corresponding values for y. Sketch the equation.
2. For each of the following graphs, write the general form of the linear equation that would represent the specific relationship.



3. Two variables that obviously increase and decrease together may not belong in the same model. Explain why.
4. Using the general rules that serve as a guide to whether a time series should be presented on an arithmetic or a ratio (logarithmic) scale, explain which scale you would choose for the following:
   a. Unemployment rate.
   b. Price levels.
   c. Income.
5. Suppose that you have constructed a time series showing how income levels in the United States have changed over the past few decades. Would there be any point in constructing a graph of the distribution as well? Would it give you any additional information?
6. Given the following list of economic variables, determine whether each is a *stock* or a *flow:*
   a. Salary.
   b. Sales.
   c. Wealth.
   d. Volume of trading on the New York Stock Exchange.

4

Demand and Supply

As you read and study this chapter, you will learn:

- the simple analysis of demand
- three main elasticities of demand: price elasticity, income elasticity, and cross elasticity
- the simple analysis of supply
- the elasticity of supply
- how supply and demand interact to determine price and quantity in the market

The main task of microeconomics is to explain the relative values of economic goods. Those values are mostly determined by forces operating in markets throughout the economy. ***The resulting market values are prices, and prices are at the core of microeconomics.***

Table 1 presents a wide variety of such prices. They are real—formed and changed incessantly in real markets. Their *relative* levels are crucial. To see the importance of relative prices, imagine that you own a one-ounce gold coin. It is a small object, barely able to cover your fingertip. Yet that little piece of metal has recently been worth $452. And, as Table 1 shows, with that one gold coin you could buy 481 pounds of beef, 4,475 pounds of potatoes, 32,873 pounds of old newspapers, or 11 airplane trips between New York and Detroit. How can such a small bit of gold buy so many other goods? There are other puzzling comparisons. For example, why is the value of 583 dozen eggs only equal to 307 pounds of coffee, 13 barrels of oil, or 20 hours of an electri-

Table 1 ***Prices of Selected Commodities and Retail Items on a Recent Day***

Commodity (unit of measurement)*	Price per Unit	Amount of the Commodity Equivalent to 1 ounce of Gold Priced at $452 per ounce
Foods		
Beef, 700–900 pound carcass (per pound)	94¢	481 pounds
Butter, grade AA (per pound)	$1.52½	296 pounds
Chicken broilers, dressed (per pound)	45⅙¢	1,001 pounds
Coffee, Brazilian (per pound)	$1.47	307 pounds
Eggs, large white (dozen)	77½¢	583 dozen
Orange juice, frozen concentrate (per pound)	$1.16	390 pounds
Pork bellies (per pound)	57¢	793 pounds
Potatoes, round white (per pound)	10.1¢	4,475 pounds
Sugar, cane, raw (per pound)	16.46¢	2,746 pounds
Grains and Feeds		
Corn, No. 2 yellow (per bushel)	$2.59½	174 bushels
Oats, No. 2 millings (per bushel)	$2.33	194 bushels
Rice, No. 2 milled (per pound)	22¢	2,055 pounds
Sunflower seed, No. 1 (per pound)	11.3¢	4,000 pounds
Wheat, No. 2 soft red (per bushel)	$4.15¾	109 bushels
Fibers and Textiles		
Cotton, 1 1/16-inch strand (per pound)	57.5¢	786 pounds
Print cloth, cotton, 48-inch (per yard)	78¢	579 pounds
Wool, fine staple (per pound)	$2.85	159 pounds
Metals		
Aluminum, ingot (per pound)	76¢	595 pounds
Copper, refined (per pound)	80¢	565 pounds
Lead (per pound)	35¢	1,291 pounds
Nickel, plating grade (per pound)	$3.50	129 pounds
Steel scrap, 1 heavy melt (per ton)	$80.00	11,300 pounds
Tin (per pound)	8.24¢	5,485 pounds
Precious Metals		
Gold (per troy ounce)	$452.00	1 ounce
Silver (per troy ounce)	$8.59	52.6 ounces
Petroleum		
Crude, Saudi Arabia light (per barrel)	$34.00	13.29 barrels
Gasoline, regular, wholesale (per gallon)	$1.00½	450 gallons
Gasoline, unleaded, wholesale (per gallon)	$1.02½	441 gallons
Miscellaneous Commodities		
Cowhides, light native (per pound)	57¢	793 pounds
Newspapers, old, No. 1 (per ton)	$27.50	32,873 pounds
Rubber, smoked sheets (per pound)	45⅜¢	995 pounds
Retail Items		
Apartment rent, 1 bedroom, unfurnished (per month)	$195.00	2.32 months
Ball-point pen	29¢	1,559 pens
Discount air fare, Detroit to New York, one-way	$39.00	11.90 trips
Motion picture ticket	$3.50	129 tickets
Unskilled labor, mininum wage (per hour)	$3.60	126 hours
Skilled labor, electrician (per hour)	$22.50	20.1 hours

Sources: *Wall Street Journal*, Department of Commerce, *Business Week* magazine.
*Most commodity prices are for delivery in New York. Some retail prices are estimates.

cian's labor? What causes nickel to be ten times as valuable as lead, and wool five times as valuable as cotton?

Note, too, that the gold has great power over human labor. That one-ounce coin will buy 126 hours of unskilled labor, which is more than three weeks of full-time work. One pound of gold will buy you nearly an entire year of full-time work by an unskilled laborer.

Table 1 is long, but it is only a small selection from thousands of commodities and retail items. Every day the market system adjusts all of these prices by numberless interactions of supply and demand. *The prices express the relative values of the goods.* Altogether they are at the heart of microeconomic activity.

Microeconomics seeks to explain those relative values, by using a relatively few powerful tools. Moreover, microeconomics shows how market prices can reconcile major social conflicts by solving two fundamental problems. First, markets *coordinate the actions of producers and consumers,* even though the two groups may never meet or communicate directly. Wheat farmers need to sell their grain. Bakers need flour to bake bread. Families need bread. The markets for wheat, flour, and bread enable them to serve one another's needs, while pursuing their own separate interests. Second, markets *resolve conflict.* Wheat farmers want high prices. Families want cheap bread. The markets that link them together resolve this antagonism in an impersonal way, without bitter face-to-face strife.

Microeconomics relies heavily on *comparative static analysis,* which works as follows: First, an equilibrium situation is defined (for example, certain supply and demand conditions for oil yield a price of $34 per barrel and a quantity of 20 million barrels per period). Then the value of one variable is changed, and the resulting effect on the outcome is traced through (a 10 percent rise in world incomes causes oil's price to rise to $36 and the flow to 22 million barrels). The method compares the first static outcome with the second; hence the name "comparative statics."

The method requires holding many complex factors constant, so that attention can be focused on single changes. Moreover, it works especially well with small *marginal* changes. Since comparative statics has such a narrow focus, it can disentangle the many forces at work in complex economic processes.

The first step in microeconomics is to explain the outcomes of individual markets. The market is the arena where demand and supply meet. ***The term* market *has a precise economic meaning: a grouping of buyers and sellers who exchange a specific good at a price.*** Many people participate, and there is competition among both buyers and sellers. A market also has rapid adjustments in prices and quantities.

Each market is described by two main dimensions: the nature of the good and the geographic area. Each market is defined so that it contains only goods among which consumers can freely substitute. For example, two brands of English muffins would be in the same market, but sledgehammers and saws are in different markets. Atari and Odyssey electronic games are not identical goods, but they would both be included in the market for such games. The second dimension of a market, *geographic area,* is also important in defining markets. Since bread is usually baked and sold locally, the Chicago bread market and the Detroit bread market are separate markets. On the other hand, since microwave ovens are transported nationwide, the market for a particular brand would be the entire United States.

Once a specific market is defined by both its product and its geographic area, then one can examine the workings of demand and supply in that particular market

setting. That is precisely what you will be doing in this chapter. The demand and supply sides of the market are presented separately. Then we show how demand and supply forces interact to determine market price and quantity.

Demand

Influences on demand

The analysis of **demand** focuses on the consumers' (or buyers') side of the market. It seeks to explain what demand is and how it affects market price and quantity. As you know from experience, many factors determine how much of a commodity consumers are willing to buy at various prices.

The *price* of the good is an obvious and important influence on purchases. Simple intuition tells you that at higher prices people will purchase less of a good than they will at lower prices, if all other things remain unchanged.

Income (or buying power) is another important influence on the quantity of a good that people will want to buy. As your income rises, you can afford to buy more, and your purchases of many goods will increase. Yet there are also goods that you would buy less of as your income rises. For example, if your own income were suddenly increased, you might buy less hamburger and more steak, or less margarine and more butter.

Preferences determine the relative importance of other variables in determining how much of each product will be bought at each price. Preferences are not always easy to explain, but economists regard them as a crucial influence on demand. The companies selling consumer goods also recognize the importance of preferences; they spend over $40 billion a year in the United States on advertising in attempts to direct consumer preferences toward their own products.

The amount of a good that consumers wish to purchase is influenced not only by the item's own price, but also by the *price of other goods*. These other related goods can be *substitutes* or *complements*. A substitute is a good that can be used in place of another commodity. For example, margarine and butter are widely recognized as substitutes for each other. If the price of butter increases, consumers may substitue margarine. Thus, the amount of margarine purchased will change because of the price change of butter. In contrast, goods that are complements are used together rather than instead of each other—for example, cars and gasoline. As the price of gasoline increases and people buy less of it, you would also expect fewer cars to be sold as consumers switch to buses, form car pools, or move closer to their jobs.

Population will also affect the amount of a good purchased. More consumers mean that more of a good will be purchased, other things being equal.

Patterns of consumption will also be affected by *income distribution*. In a country with a very even distribution of income, there will be less demand for certain luxury goods than there would be in an economy with the same total income divided very unequally among the very rich and the very poor.

Finally, *expectations about future prices* can affect purchases. If consumers expect the price of a good to rise sharply in the near future, they may buy more of the good now to avoid paying higher prices later. If consumers expect a price to fall, they are likely to postpone purchases.

All of these conditions can influence the quantity of each good that consumers will purchase. The economist's task is to disentangle all of these different factors. How does microeconomics isolate the effect of just one of these factors on the quantity of a good that consumers will want to purchase?

The procedure is fairly simple. First, hold all the influences constant at a given level and measure the quantity that consumers will want to buy. Then allow one influence to vary, while still holding the other influences constant. For example, hold income, population, and all of the other influences on demand constant, and only vary the price. Since price is the only influence on demand that is allowed to vary, any changes in quantity must be caused by the changes in price. Thus, you can define and isolate the specific relation between price and quantity. Note that the other factors that affect demand are not ignored; their influence is simply held steady.

You can hold all factors constant and then vary any one influence. If you vary income, you are isolating the income-quantity relation. If you vary population, you are identifying the population-quantity relation. If you allow price to vary to bring out the price-quantity relation, then you can derive an important tool of economic analysis, the *demand curve.*

The demand curve

The demand curve for any good relates the quantities of the good that consumers wish to buy to the *price* of the good. The underlying logic in deriving a demand schedule can be summarized as shown below. All influences except price are held constant. Then quantity varies only as price varies, so that the demand curve reveals the relationship between price and quantity.

Table 2 Conditions for a Simple Demand Curve

Price	Quantity (gallons per week)
$3.50	6,200
3.00	8,000
2.50	10,000
2.00	12,300
1.50	15,000
1.00	18,000
.50	21,500

A demand curve can represent the price-quantity relation either for an individual consumer or for an entire market. This chapter concentrates on *market* or total demand curves, which come from adding all of the individual demand curves for a particular good.

The market demand curve can best be shown by an example. A demand schedule for gasoline is presented in Table 2, showing how many gallons of gasoline consumers would be willing to buy at various prices. For example, at a price of $1.00 per gallon, consumers would be willing to buy 18,000 gallons of gas per week. From your experience as a consumer, what is likely to happen if the market price of gasoline increases? At a higher price, people will probably want to buy less gasoline. This *inverse relation between price and quantity* makes intuitive sense: ***At higher prices, people will want to buy less of a good; at lower prices, they would be willing to buy more.*** The figures in Table 2 illustrate precisely that.

Influences on Demand

The quantity of a specific good that consumers will buy — *depends on* — The price of that good; Consumers' income levels; The number of consumers; Income distribution; Consumers' preferences; The price of related goods; Expectations about future prices

The Demand Curve Relates These Two Variables

These Influences are Held Constant

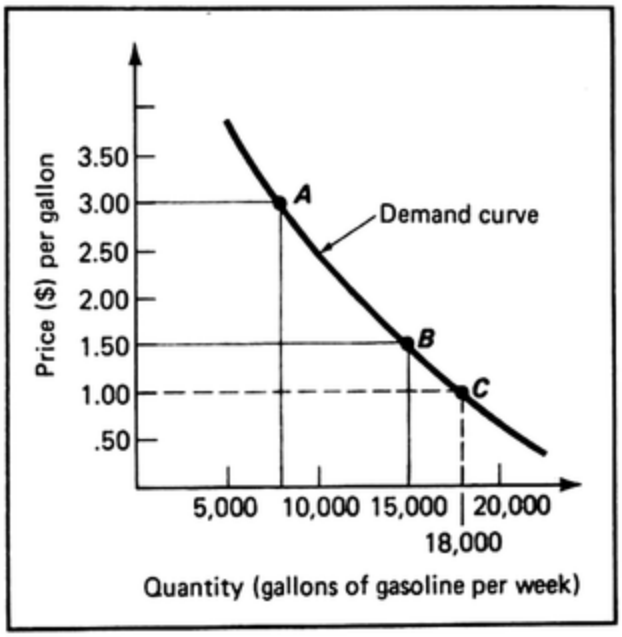


Figure 1 The market demand curve relates price and quantity

At the price of $1.50 per gallon, the consumers in the market wish to purchase 15,000 gallons per week. At $3.00 per gallon, they want only 8,000 gallons per week. Graphing other quantities, at other prices, traces out the demand curve.

The inverse relation between price and quantity becomes even clearer if the figures in Table 2 are graphed, as in Figure 1. The demand curve showing the relation between price and quantity has a negative or downward slope, because as price rises, consumers want to buy less of the good.

Although the downward slope of the demand curve probably seems obvious, you should be able to explain the inverse relation between price and quantity in more precise terms than intuition. There are two major reasons for the downward slope:

1. ***Substitution effect.*** As the price of good increases, the substitutes for it (with their now relatively lower prices) look more attractive. Consumers are therefore likely to buy more of the substitutes and less of the good whose price has gone up.
2. ***Income effect.*** As the price of a good increases, the consumers' purchasing power decreases. Consumers can afford to buy less of all goods, including the one whose price has risen. If the price of razor blades or a bag of pretzels increases, the income effect will be insignificant. But for goods like housing or heating oil that account for a large percentage of people's budgets, the income effect may be substantial. For example, suppose that dormitory rental rates and apartment rents increase considerably. Since housing costs account for so much of your student budget, you will probably have to consume less housing space by rooming with more people than you had originally expected to room with. Instead of two roommates in your apartment, you may have four. Furthermore, to pay the rent you will probably also have to reduce your consumption of *other* things, such as new clothes and entertainment.

For these two reasons, then, your intuition that there is an inverse relation between price and quantity is sound.

Several other points about the demand curve should also be kept firmly in mind. First, the time period for demand must be exactly specified. To know that consumers want to buy 15,000 gallons of gasoline at $1.50 is not very helpful, unless one knows whether that is 15,000 gallons a day, week, month, or year. Without that information, you cannot define demand in any meaningful way. Second, the various quantities expressed in Table 2 and Figure 1 are alternative *desired* quantities, at alternative prices. They represent the amounts that consumers would like to buy at the various prices, and not the amounts they actually succeed in purchasing. For example, at $2.00 a gallon, consumers might want to buy 12,300 gallons of gas *if they could.* But suppose that 12,300 gallons

of gas are not available. Actual purchases would then fall short of 12,300 gallons, but that does not alter the demand curve. The curve expresses consumer preferences, not necessarily actual market outcomes. In short, the demand curve is the relationship between price and desired purchases, which may not actually occur.

Finally, you can think of a demand curve in two different ways. Consider the demand curve in Figure 1. (1) The curve shows the maximum quantity of the good people want to buy at a given price—15,000 gallons of gas at $1.50 per gallon. (2) The curve also shows the highest price that consumers will pay to get a given quantity of the good—$1.50 per gallon for 15,000 gallons. People would rather be under the curve, paying less for each given quantity, but the curve itself shows the *maximum* that they are willing to pay.

Because it is diagrammed as simply a line, the demand curve may seem trivial. But, in fact, it summarizes very important information. It shows how much money people are willing to sacrifice to get a good. Of course, you could try to discover preferences by asking all consumers detailed questions, but that would be extremely difficult and give doubtful results. Meanwhile, every day in thousands of markets, consumers put their preferences on the line by making practical choices and payments. These actual and expressed preferences are embodied in the demand curve, making demand an important and powerful concept. To use demand curves correctly, though, the distinction between *demand* and *quantity demanded* must be drawn carefully and precisely.

Quantity demanded and demand

Quantity demanded *refers to a specific price-quantity combination: a particular point on a demand schedule.* In Table 2, 15,000 gallons would be the quantity demanded at $1.50. This would correspond to Point *B* on the demand curve in Figure 1.

Demand refers to the entire price-quantity relationship: to the whole demand curve in Figure 1, or the entire column of figures in Table 2.

The difference between quantity demanded and demand is crucial in analyzing changes in a particular market. Since quantity demanded refers to a particular point on a demand curve, *a change in quantity demanded refers to a movement along the demand curve from one price-quantity combination to another.* In Figure 1, the movement from Point *B* (15,000 gallons at $1.50) to Point *A* (8,000 gallons at $3.00) would be a change in quantity demanded. Along a given demand curve, the only change that can cause a change in quantity demanded or a movement along the demand curve is a change in price.

Since demand refers to the entire price-quantity relation, the phrase *"a change in demand" refers to a shift in the entire demand curve.* Such a shift in demand can be caused by a change in any influence on quantity *except price.* Take Figure 2 as an example. Curve I is the original demand curve. Now suppose that population increases, all other things remaining unchanged. Because there are now more consumers, a larger quantity of the good would be demanded at every price. The entire demand curve shifts upward and to the right, illustrated by the shift from Demand Curve I to Demand Curve II. Now at a price of P_a, consumers wish to purchase the higher quantity of Q_b.

If population decreases, or if a change in preferences makes consumers like the good less, then consumers would demand less of the good at every price. This decrease in demand would be represented by a leftward shift of the demand curve. But remember that a change in price cannot by itself cause a change in demand. The result of a price change is a *movement along the*

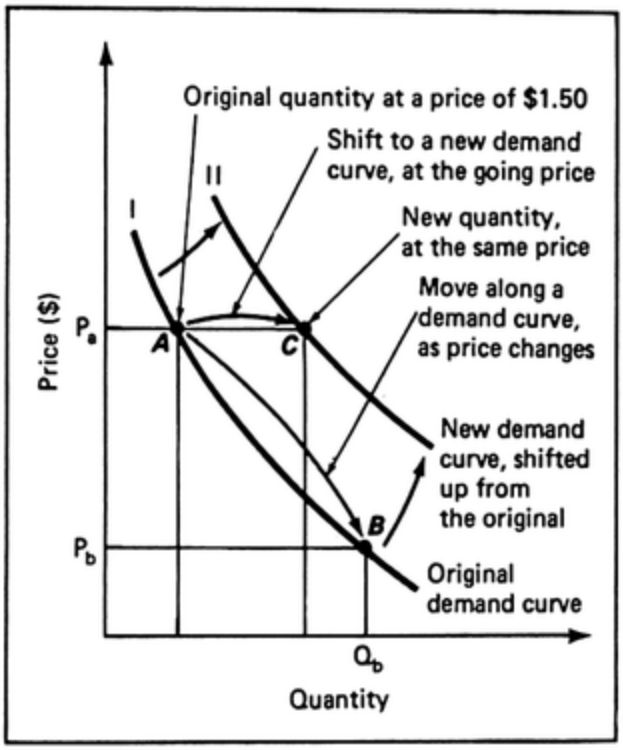


Figure 2 The distinction between a change in demand and a change in quantity demanded

If price falls from P_a to P_b, consumers would move along the demand curve from Point *A* to Point *B*. If any other influence on quantity changes, the demand curve will shift, and a different quantity of the good will be demanded at any given price.

demand curve, not a shift in the entire schedule. Keeping changes in quantity demanded (movements along a demand curve) separate from changes in demand (shifts in demand curves) is crucial when you use demand to analyze changes in markets.

By now, you should have a fairly good idea of how to work with demand curves. You know what causes shifts of these curves and what causes movements along demand curves, and you know why demand curves have a downward or negative slope. But while the general downward-sloping form of demand curves has been discussed, one issue that has not yet been dealt with is the exact *shape* of demand curves, which depends upon how responsive quantity is to price changes.

Consider the two demand curves in Figure 3. Curve I is relatively steep, while Curve II is relatively flat. If the price increases from $1.50 to $2.00, both curves show a decrease in quantity demanded. In the case of the steeper demand curve, Curve I, the drop in quantity from the 50-cent increase in price is 1,500 gallons per week, from 15,000 to 13,500 gallons. For the flatter Demand Curve II, the 50-cent rise in price causes a larger drop in quantity of 5,000 gallons per week, from 15,000 to 10,000 gallons. In other words, quantity

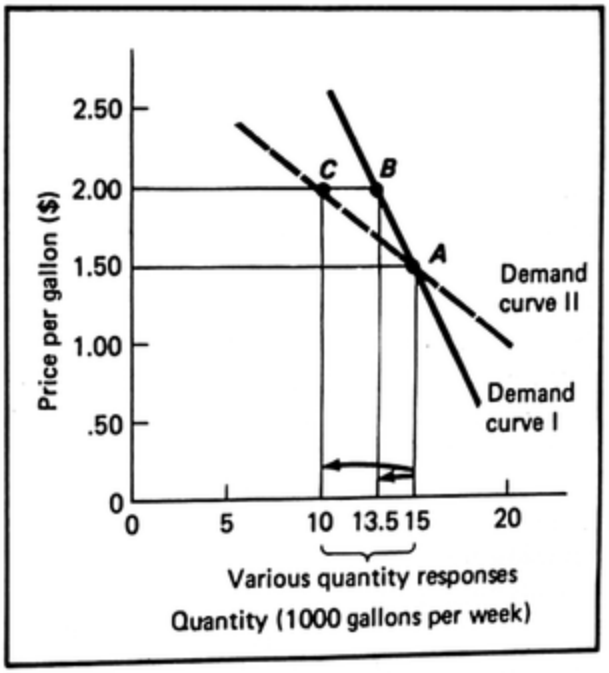


Figure 3 The responsiveness of quantity demanded to changes in price

Both Demand Curve I and Demand Curve II show quantity demanded falling when price increases, and rising when price decreases. Yet the degree of responsiveness of quantity demanded to a given price change is different for each curve. For example, suppose that the price increases from $1.50 to $2.00. Along the steeper demand curve, Demand Curve I, the quantity demanded falls from 15,000 to 13,500 gallons, a decrease of 1,500 gallons. Along the flatter demand curve, Demand Curve II, the quantity demanded falls from 15,000 to 10,000 gallons, a decrease of 5,000 gallons. *For the given price change* from $1.50 to $2.00, the quantity demanded represented by Demand Curve II is more responsive to price changes than the quantity demanded represented by Demand Curve I.

demanded is less responsive to price changes along the *AB* segment of Demand Curve I than along the *AC* portion of Demand Curve II.

How responsive quantity is to price changes is of great importance to any producer or supplier of a good. The reason lies in the influence this responsiveness has on total expenditure or revenue. Total revenue can be expressed as:

$$\text{Total Revenue} = \text{Price per unit} \times \text{Number of units sold}$$

or

$$\text{Total Revenue} = \text{Price} \times \text{Quantity}.$$

Suppose that a producer decides to raise prices to increase the total revenue taken in from the sale of a good. The plan may not work. You already know that as price increases, quantity demanded decreases. But while the increased price would, in itself, cause the firm's total revenue to increase, the resulting decrease in quantity would exert an opposite or downward pull on the total revenue gained from consumers' spending. Total revenue may therefore increase or decrease as a result of the price increase. Figure 4 illustrates what may happen to total revenue as price increases. If the price is increased from \$2 to \$5, the quantity demanded drops from 500 units to 300 units. Total revenue *increases* as a result of this price increase, from \$2 × 500 = \$1,000 to \$5 × 500 = \$1,500. But if price is increased further, from \$5 to \$7, total revenue *decreases* from \$5 × \$300 = \$1,500, to \$7 × 100 = \$700.

Whether total revenue will increase or decrease when price changes depends on which change is larger—the change in price or the change in quantity. To measure the relative sizes of price and quantity changes, economists use the concept of price elasticity of demand.

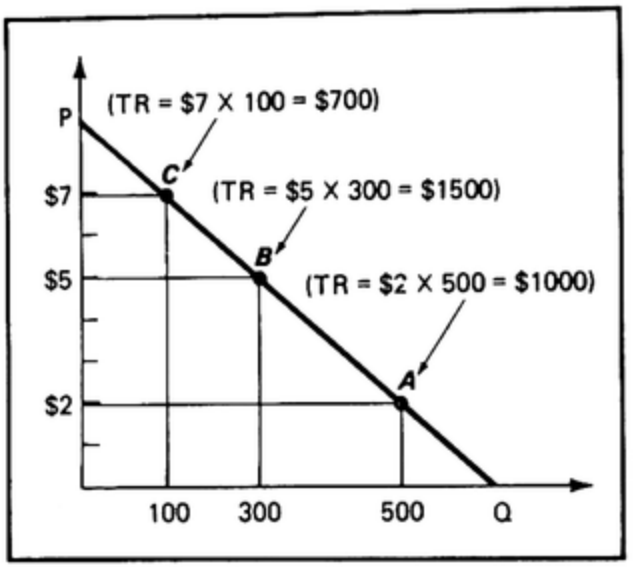


Figure 4 Total revenue at different points on the demand curve

Total revenue may either rise or fall as price increases.

Price elasticity of demand

Price elasticity of demand ***measures the relative responsiveness of quantity demanded to a change in price.*** The formula for elasticity is:

$$\text{Price elasticity} = \frac{\text{percentage change in quantity}}{\text{percentage change in price}} \text{ or } \frac{\%\Delta Q}{\%\Delta P}.$$

Any elasticity is simply a ratio between a cause and an effect, always in percentage terms. The cause goes in the bottom half or denominator of the ratio, while the effect is in the top half or numerator. For price elasticity, assume that the price change is the cause, and the change in quantity is the effect.

When you calculate elasticities, it is important to work with percentages to avoid the problem of scale. After all, it would hardly be fair to declare that the quantity demanded of gasoline is more responsive to changes in price than automobiles, because a \$1.00 change in the price of each good called forth a bigger change in gasoline sales than it did in car sales. By

using percentages, the $1.00 price change is put in perspective.

Note that since price and quantity are inversely related, always moving in opposite directions, price elasticity will always be a negative number. The custom is to *ignore the negative sign and discuss only the absolute level* of the price elasticity.

An example can clarify both what elasticity means and how to measure it. Suppose that the price of gasoline rises from $1.50 to $2.00 a gallon in your town, while all other influences on the quantity of gasoline demanded remain unchanged. This causes the local motorists to cut their use from 15,000 to 13,500 gallons per week. The change is shown in Figure 3 along Demand Curve I, represented by a movement from Point *A* to Point *B*. Price elasticity can then be calculated as:

$$\text{Elasticity} = \frac{\text{effect}}{\text{cause}} = \frac{\%\text{ change in } Q}{\%\text{ change in } P}$$

$$= \frac{-1{,}500/15{,}000}{+\$.50/\$1.50} = \frac{-10\%}{+33.3\%} = .30.$$

This method gives the elasticity over the range between Points *A* and *B*, using *A* as the reference or starting point. Unfortunately, this raises a technical problem, for if Point *B*'s values were used instead as the reference in calculating the percentages, the elasticity figure would differ:

$$\text{Elasticity} = \frac{-1{,}500/13{,}500}{+.50/\$2.00} = \frac{11.1\%}{25\%} = .444.$$

The difference in the estimates would be even larger if price doubled to $3.00. The rise would be *100 percent*, but an exact reversal to $1.50 would be only a cut of *50 percent* of $3.00.

To avoid this problem, the midpoint of the range is commonly used as the base. In the present case:

$$\text{Elasticity} = \frac{-1{,}500/14{,}250}{+.50/\$1.75} = \frac{10.53}{28.57} = .367.$$

This gives the average elasticity for that segment of the curve. Another approach would be to measure tiny changes (e.g., a price rise from $1.50 to $1.52), but that would be difficult.

Price elasticity of demand may vary between zero and infinity. With a zero value, demand is *perfectly inelastic*, as shown in Panel I of Figure 5. In this case, quantity demanded does not respond at all to price changes. At the other extreme, where elasticity is infinite, demand is said to be *perfectly elastic*, as shown in Panel II of Figure 5. In this case, people will demand an indefinitely large quantity of the good at the present price, but a zero quantity if the price rises even slightly. Between these extreme cases lie most actual demand situations: Quantity usually responds in some finite degree to price changes, such as with the demand schedule in Panel III.

Make sure that you do not confuse the *slope* of a demand curve with elasticity. The slope of a demand curve between two points would be $(P_2 - P_1)/(Q_2 - Q_1)$. Its elasticity for this same segment would be $\frac{Q_2 - Q_1}{(Q_2 + Q_1)/2} / \frac{P_2 - P_1}{(P_2 + P_1)/2}$, using the midpoint method. A straight-line demand curve, as illustrated in Panel III of Figure 5, will have a constant slope, but *there will be differing elasticities at each point.* This is shown in Figure 6. Generally, at the upper end of a straight-line demand curve, the elasticities are greater than at the lower end.

Aside from the zero and infinite elasticities represented in Panels I and II of Figure 5, there are few exceptions to this rule of differing elasticities along a demand curve. One such exception is the rectangular hyperbola illustrated in Panel IV of Figure 5. At each point, price times quantity yields a total expenditure that is the same as at every other point. This means that price changes must always be

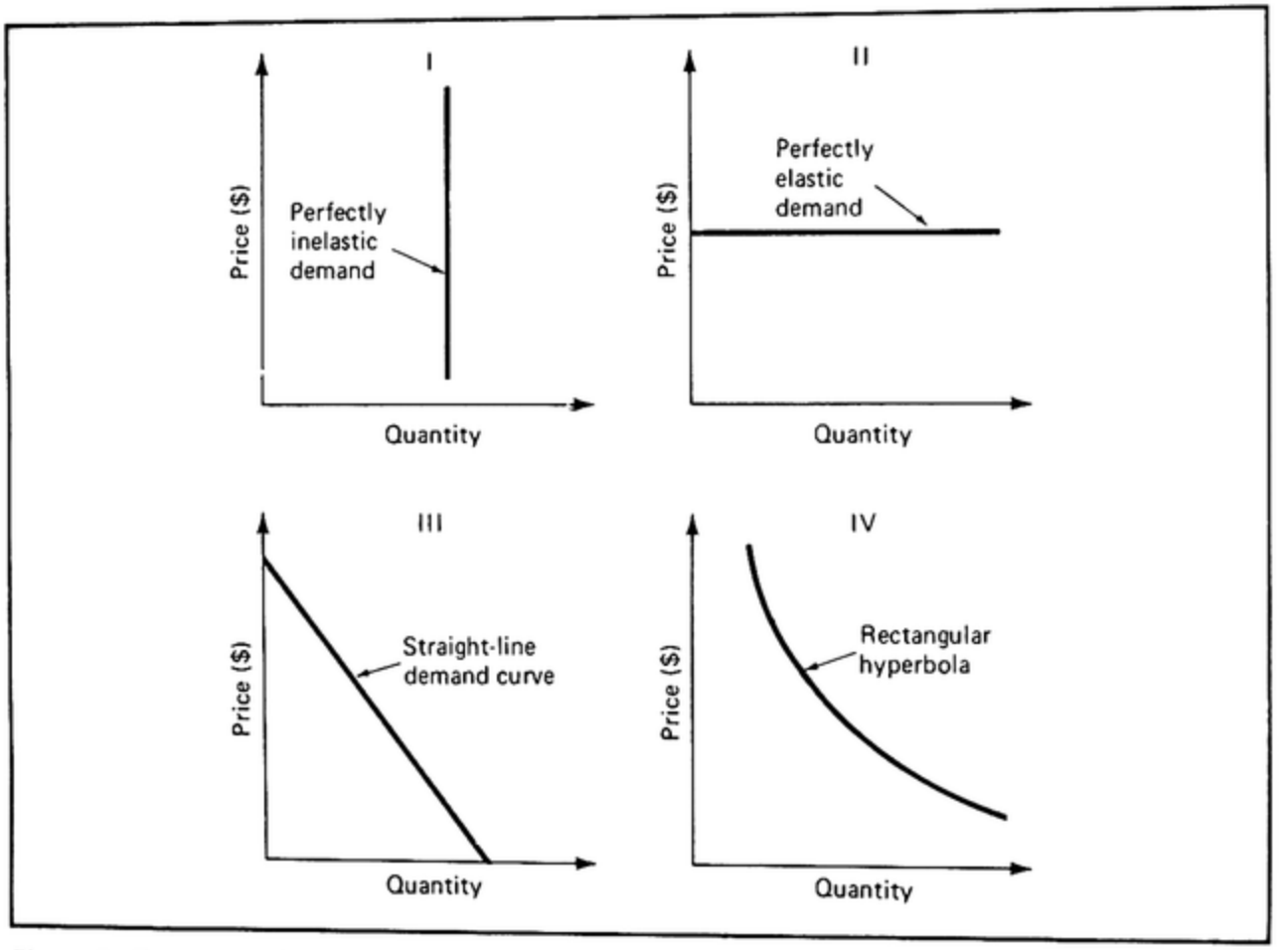


Figure 5 Four examples of demand curves: Perfectly inelastic, perfectly elastic, a straight line, and a rectangular hyperbola

The vertical demand curve (I) is perfectly inelastic. No price change is large enough to change the quantity demanded. The horizontal demand curve (II) is perfectly elastic. At that price, people are willing to consume any amount from zero to infinity. Even a slight price increase will cut their purchases to zero. The straight-line demand curve shown here (III) touches both axes. Elasticity varies along it. The hyperbolic demand curve (IV) has a constant elasticity of 1 throughout.

met by equal and offsetting quantity changes. Therefore, the curve has a constant elasticity at every point.

Once you have learned to calculate an elasticity, you still have to interpret its value. What should you make of an elasticity of 0.088, or 0.3, or 3.0, or 9.67? What do these elasticities tell you?

Elasticities are divided into three simple categories, according to which change dominates: the percentage change in quantity or the percentage change in price. If elasticity is greater than 1 (remember, by convention, the minus sign is ignored), then the percentage change in quantity (the numerator) must be greater than the percentage change in price (the denominator). For example, a 10 percent rise in the price of hang gliders causes a 20 percent fall in the number bought: elasticity is 2.0. This case is called *elastic demand*, in which quantity demanded is considered *relatively responsive* to changes in price. If elasticity is less than 1, then the percentage change in quantity (the numerator) must be less than the percentage change in price (the denominator). This is a case of *inelastic demand*, in which quantity de-

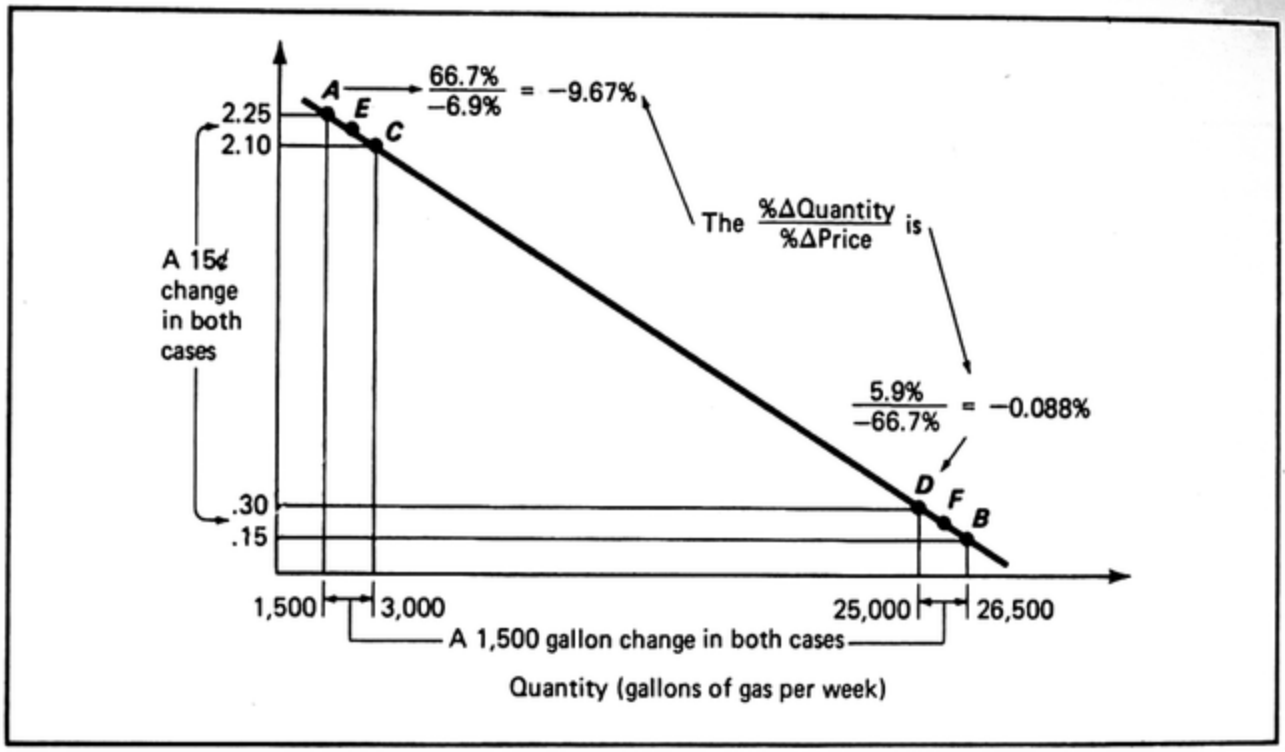


Figure 6 Elasticity varies along a straight-line demand curve

Segments *A–C* and *B–D* have the same slope, but the elasticities—which involve different percentage changes—are sharply different. They are calculated here using the midpoint formula (the two midpoints are *E* and *F*).

manded is considered relatively unresponsive to changes in price. Finally, cases in which elasticity just equals 1, meaning the percentage change in quantity equals the percentage change in price, are considered to be borderline cases, referred to as *unitary elasticity*.

Elasticity and total revenue

Once you know the price elasticity of demand, you can predict how total revenue will change when price changes. Remember that total revenue equals price times quantity. When price changes, all other influences remaining constant, quantity will move in the opposite direction. Total revenue, pulled in two different directions, will move toward the dominant change. This is because price elasticity tells which change is larger, the change in quantity or the change in price. Knowing that, you can predict in which direction total revenue will move.

If demand is *elastic*, then the quantity change dominates, and so total revenue will move in the same direction as quantity. If demand is inelastic, then the price change dominates the change in quantity, and total revenue will move in the same direction as price. If the elasticity is unitary or 1, then the percentage changes in price and quantity are equal, and there is no change in total revenue. The relation between elasticity and total revenue can be summarized as:

Elasticity ($\%\Delta Q/\%\Delta P$)	Direction of Change in P	Direction of Change in Q	Direction of Change in TR
Elastic demand (Elasticity > 1; $\%\Delta Q > \%\Delta P$)	increase	decrease	decrease
	decrease	increase	increase
Inelastic demand (Elasticity < 1; $\%\Delta Q < \%\Delta P$)	increase	decrease	increase
	decrease	increase	decrease
Unitary demand (Elasticity = 1; $\%\Delta Q = \%\Delta P$)	increase	decrease	no change
	decrease	increase	no change

Determinants of the price elasticity of demand

You have seen how to measure and interpret elasticity, but an important question still remains: What influences the price elasticity of demand? Why, in other words, is the demand for some goods very elastic or responsive to price, while the demand for other goods is very inelastic, changing only slightly when price changes? The following four influences are usually regarded as being the most important in determining each good's elasticity of demand:

1. *Necessity: Certain elementary things are necessary for life, such as food, water, and shelter.* The more you need such goods, the harder it is to cut back on purchases when the price increases. This would make demand for the good relatively inelastic. Other examples of necessities are certain medicines, textbooks, a bed, and heating fuel. Of course, the degree of necessity for some items can vary from one person to the next. Coffee may be an absolute necessity to you, but an optional good to someone else.

2. *The number of available substitutes.* If a good has many close substitutes, it is easy to cut back on purchases in response to a price change. This would make the demand for the good relatively elastic. This is why demand for a specific type or brand of a good is likely to be more elastic than demand for the general category of that good. For example, demand for a new Ford or Dodge will be more elastic than demand for new cars in general.

3. *Time.* As prices change, consumers plan to adjust the quantity they purchase, but this takes time. Plans must be changed; other purchases must be altered. The longer the interval, the greater the changes can be. Therefore, demand over longer time periods tends to be more elastic. For example, suppose that the price of gasoline rises from $1.50 to $3.50 a gallon. Consumers may immediately try to decrease their consumption of gas, but their opportunities for immediate changes are limited. After a period long enough for people to begin to rely on public transport, form car pools, move closer to work, buy bikes, or trade in gas-guzzling cars for more fuel-efficient models, you would expect to see a larger adjustment in quantity: that is, a more elastic demand.

4. *The percentage of income spent on a good.* If you spend a large amount of your income on a particular good, then you are forced to adjust when the price increases. This would tend to make the demand for that good relatively more elastic. For example, dormitory fees or apartment rentals probably account for much of your student budget. If those fees or rentals increase, you may be forced to buy less living space by sharing your apartment with more students. On the other hand, if the price of cole slaw increases, you hardly have to make major adjustments in your purchases. The price increase is simply not significant.

Income elasticity of demand

The income elasticity of demand *measures the responsiveness of quantity demanded to changes in income.* The formula for income elasticity is:

$$\text{Income elasticity} = \frac{\text{percentage change in quantity demanded}}{\text{percentage change in income}} = \frac{\%\Delta Q}{\%\Delta \text{Income}}.$$

In this case, all influences on demand other than income are held constant. As income varies, the relation between quantity demanded and income can be determined.

To calculate income elasticity, consider a town in which family income averages $15,000 per year. Average family purchases each year are 40 pounds of

Estimate an Elasticity Cautiously

It is often possible to estimate elasticities from actual price and quantity changes. But when other factors are changing too, the estimates are hazardous.

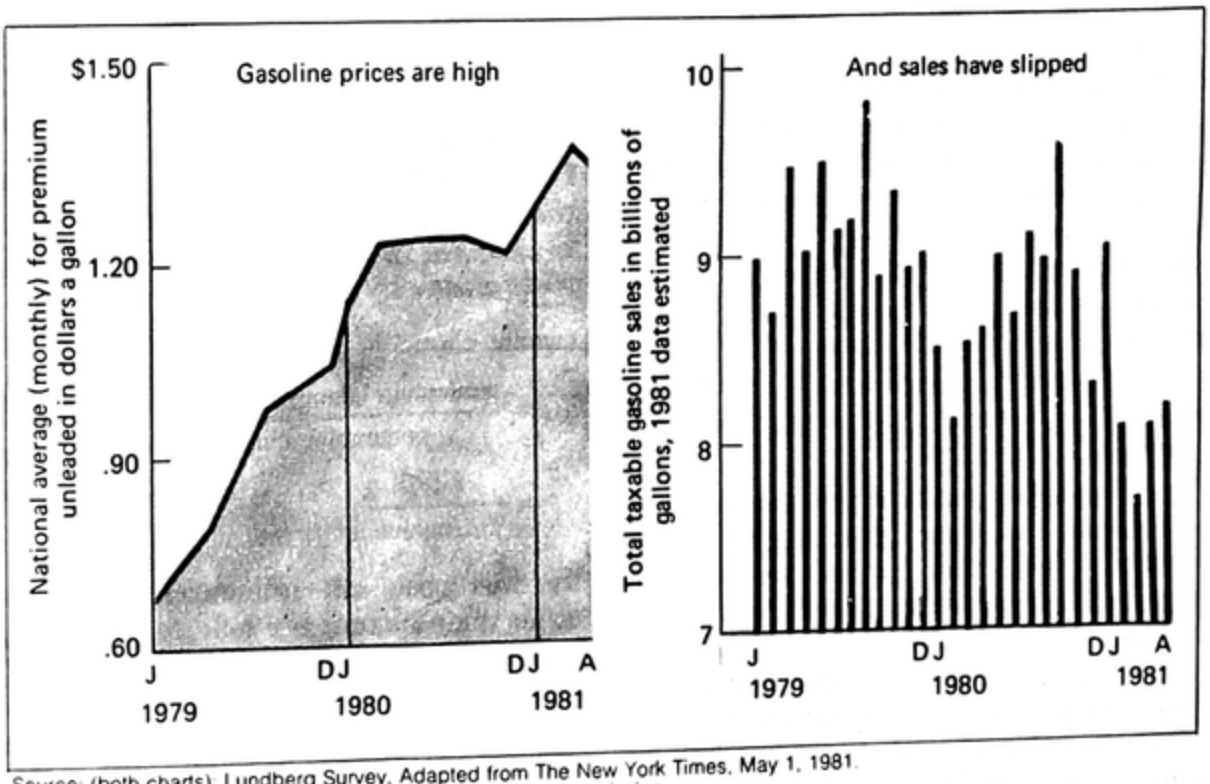
For example, a recent *New York Times* news story included adjacent charts of gasoline prices and quantities. They showed that while gasoline prices rose in the United States from 70 cents to about $1.30 per gallon during 1979–1981, the quantity of gasoline sold fell from about 9.4 to about 8.0 billion gallons per month. Not only do the directions of the two changes fit basic logic, but they also permit one to calculate a specific value for elasticity as follows:

The price rise was 60 cents per gallon. That is a 60 percent rise, calculated using the midpoint of the 70-cents-to-$1.30 range, which is $1.00. Meanwhile, the flow of gallons declined by 1.4 billion per month, which is 16 percent of 8.7 billion gallons (the midpoint of the 9.4-to-8.0-billion range). The elasticity of demand appears to be

$$\frac{\%\text{ change in gallons}}{\%\text{ change in price}} = \frac{-16\%}{+60\%} = -0.27 = 0.27.$$

(The minus sign is ignored, of course.) Evidently the simple logic of elasticity can be applied to real data.

But there are two main cautions. First, other factors have certainly also changed during this period, so that the



Source: (both charts) Lundberg Survey. Adapted from The New York Times, May 1, 1981.
© 1981 by The New York Times Company. Reprinted by permission.

pure logic of the elasticity—to hold everything constant except price—has been violated to some degree. For example, consumers' incomes changed, and the prices of automobiles rose. Therefore, the "elasticity" number reflects more than just the change in gasoline's price. Perhaps if the other changes were small or mutually offsetting, then the calculated elasticity might not be far from the true value. Moreover, only the general magnitude of the elasticity matters (such as "about .3" or "probably between .20 and .35"). Even when the data seem to be precise, there can be errors, distortions, or unexpected factors. Therefore, the careful analyst usually speaks of approximate, not highly exact, values.

Incidentally, note that the *Times* artist cut the vertical axes dramatically on both diagrams, to make the changes more sharply visible. That is risky practice.

margarine, 5 pounds of butter, and 64 loaves of bread. Suppose that family income rises by an average of $1,500 to $16,500. As a result of this change in income, the average family now consumes 36 pounds of margarine per year, 7 pounds of butter, and 65 loaves of bread. The income elasticities of demand for these goods would be (using the midpoint method):

$$\text{Income elasticity of margarine} = \frac{\dfrac{36-40}{38}}{\dfrac{\$16{,}500-\$15{,}000}{\$15{,}750}} = \frac{-10.52\%}{+9.52\%} = -1.11$$

$$\text{Income elasticity of butter} = \frac{\dfrac{7-5}{6}}{\dfrac{\$16{,}500-\$15{,}000}{\$15{,}750}} = \frac{+33.33\%}{+9.52\%} = 3.50$$

$$\text{Income elasticity of bread} = \frac{\dfrac{65-64}{64.5}}{\dfrac{\$16{,}500-\$15{,}000}{\$15{,}750}} = \frac{+1.55\%}{+9.52\%} = +.16.$$

Do *not* ignore the sign of income elasticities. Unlike price elasticity where the sign is always negative, income elasticity may have either a positive or a negative sign. That sign provides an important clue to the nature of the good.

If the quantity of a good purchased moves in the same direction as income—increasing as income increases and decreasing as income decreases—then income elasticity will be positive. Goods with a positive income elasticity are called *normal goods*. Consumers buy more of these goods when they can afford to. In the earlier example, both butter and bread were normal goods.

If the quantity purchased of a good moves in the opposite direction from changes in income—decreasing as income increases and increasing as income decreases—then income elasticity will be negative. Goods with a negative income elasticity are called *inferior goods*. Consumers buy less of these goods as their income increases because they are substituting more desirable (but more expensive) alternatives. In the preceding example, margarine was an inferior good. As income rose, consumers switched from margarine to butter.

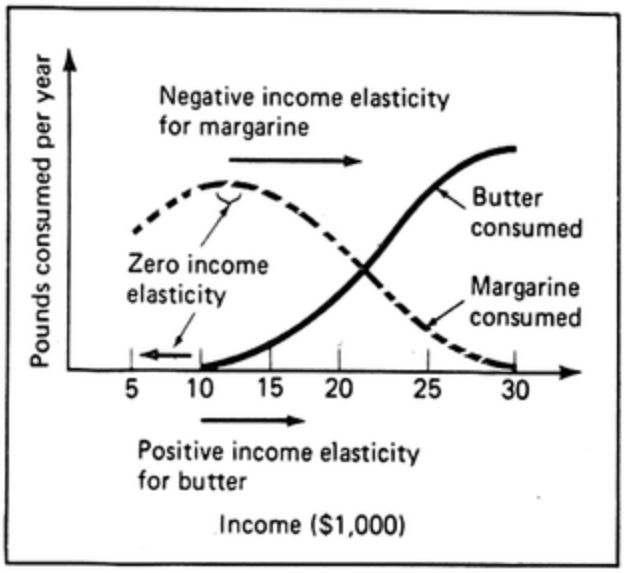


Figure 7 Income-quantity patterns for margarine and butter

At between $5,000 and $10,000 income per year, consumers purchase more margarine as their income increases. It is a normal good. For this same income range, butter shows a zero income elasticity: increases in income do not cause any increases in the quantity of butter demanded. At incomes over $10,000 per year, consumers will switch from margarine to butter. Now margarine becomes an inferior good, as increases in income cause the quantity demanded to fall. Butter becomes a normal good, with increases in income causing the quantity demanded to rise.

Income elasticity can vary with income levels. A good may be a normal good for most consumers up to a certain level of income, but become inferior at higher income levels. Figure 7 shows this for both butter and margarine. Margarine is a normal good until income rises above the $10,000 mark. Then it becomes an inferior good. Butter has a zero income elasticity up to $10,000. Consumers can't afford to buy any butter at all. At income levels above $10,000, butter becomes a normal good.

Poor families may consider potatoes and cheap cuts of meat to be normal goods. For a while, they are delighted to buy more of these goods as their income rises. But at still higher levels of income, they begin to consume less of these goods, switching perhaps to French bread and fancy steaks. Potatoes and cheap meat, once normal goods for these families, have become inferior goods at the new and higher income level.

Within the broad categories of normal and inferior goods, economists use the exact values of income elasticity to determine just how responsive quantity is to income. If the value of income elasticity is greater than 1, then the good is said to be a normal good with an *income-elastic demand.* If the value is less than 1 (but not negative), then the good is said to be a normal good with an *income-inelastic demand.* A value of 1 indicates a good with unitary income elasticity. A value of zero shows that quantity demanded does not respond at all to changes in income. And a negative value indicates an inferior good. In the earlier examples of income elasticities, butter, with an income elasticity of +3.50, was a normal good with an income-elastic demand. Bread, with an income elasticity of +.16, was a normal good with an income-inelastic demand. Margarine, with an elasticity of −1.11, was an inferior good.

Cross-elasticity of demand

Cross-elasticity of demand *measures how responsive one good's quantity demanded is to changes in the price of another good.* A pair of goods may be related because they are *substitutes* for each other (margarine and butter), or because they are used together as *complements* (gasoline and cars). The formula for cross-elasticity is:

$$\text{Cross-elasticity of demand for Good } A = \frac{\%\text{ of change in quantity of Good } A}{\%\text{ of change in price of Good } B}$$

$$= \frac{\%\Delta Q_A}{\%\Delta P_B}.$$

First, consider two goods that are *substitutes,* such as Chevrolets and Fords. If the price of Chevrolets rises, while all else

(including Ford prices) stays the same, some people who would have bought Chevrolets will now buy Fords. Thus, an increase in the price of Chevrolets would cause an increase in the quantity of Ford cars purchased. With the price of Chevrolets and the quantity purchased of Fords changing in the same direction (both increasing), the cross-elasticity will be *positive.* A *positive cross-elasticity indicates that two goods are substitutes.* The higher the value of the cross-elasticity, the closer the degree of substitution. For example, the cross-elasticity between Chevrolets and Fords is certain to be higher than the cross-elasticity between Chevrolets and bicycles.

Now consider two goods that are *complements,* such as automobiles and gasoline. If the price of gasoline increases, people will purchase both less gas and fewer of the goods used with gas, such as autos and tires. Because the increase in the price of gas causes a decrease in the quantity of cars purchased, the cross-elasticity will be negative. For example, the values might be:

$$\text{Cross-elasticity} = \frac{\%\Delta Q \text{ of cars}}{\%\Delta P \text{ of gasoline}} = \frac{-40\%}{+100\%} = -.4.$$

The negative cross-elasticity shows that the goods are complements. A negative value larger than -0.1 or -0.2 would show that they are close complements.

Close complements are not numerous. If you can think of even a few examples, you are doing well. In contrast, close substitutes are common. You should be able to think of a substitute for nearly any good or service.

Up to this point, the chapter has concentrated on the demand side of the market, analyzing how consumers' choices are expressed in demand. To complete the analysis of market quantities and prices, the next section deals with producers' choices on the *supply* side of markets.

Supply

The analysis of supply focuses on the producers' (or sellers') side of the market. Issues dealing with supply concentrate on what determines the amount of a good that producers are willing to offer. Since willingness to supply is based on several considerations, the logical starting place for an analysis of supply is with an examination of the factors that influence it.

Influences on supply

Price is one of the major influences on the quantity of a good that a producer is willing to offer. At higher prices, a producer will usually supply higher quantities of a good because, other things being equal, the higher price makes it more profitable to do so. The incentive to produce is directly related to the level of the price.

The *costs* of producing the good also influence the amount that producers will offer. Costs, in turn, are determined by two major factors:

1. *The prices of inputs.* Given the state of technology, the prices of inputs determine the cost of production. If input prices rise (e.g., when wages rise), then cost will rise and producers may have to receive a higher price to cover their costs. A fall of input prices will reduce costs and permit supply at a lower price.
2. *Technology.* Technology refers to the known ways in which inputs can be combined to produce a given output. If technology improves, so that the output can be produced using fewer inputs, then costs will be lower. That en-

ables the producers to supply each amount of the good at a lower price.

Thus, either a decrease in input prices or an improvement in technology will cause production costs to fall. The producer would then be willing to offer any given quantity of the good at lower prices.

The *goals of a firm* also influence the amount of a good that producers are willing to offer. A firm aiming to maximize its profits may offer less of a good at a given price than a firm that is trying to increase its share of market sales.

The *number of firms* offering the good will also influence the total amount of the good that producers are willing to supply, just as population affects the demand for a good. This is because the total market supply comes from summing the supply of all firms. The more firms there are, the greater will be the total quantity offered.

Changes in *the prices of related goods* also influence the amount of a good that producers will offer for sale. On the supply side, two goods are related if they use a common input. For example, a farmer may produce a variety of crops, including soybeans. If the market price of soybeans increased, all other conditions remaining the same, the farmer would find it profitable to reallocate some land from other crops to soybean production because changes in relative prices affect relative profitability. When the price of soybeans increases relative to the price of the other crops, it becomes more profitable to produce soybeans. The farmer cuts back on other crops to produce more soybeans.

Finally, *expectations about future prices* can affect producers' decisions as well as consumers' decisions. For example, if the price of a product decreases, and producers expect the price decline to be temporary, they may make only slight alterations in output. But if they see this decrease as the beginning of a long-term downward trend in prices, producers will begin a much more substantial cutback. Even if there is no current change in prices, producers may increase or decrease output on the basis of long-term predictions of the direction in which the industry is moving.

All of these influences help to determine the quantity that producers will offer. To isolate and study the influence of one specific factor on quantity offered for sale, apply the same procedure as for demand. Hold all influences constant, and then allow only one influence to change. If price is the variable that is held constant, this procedure isolates the price–quantity supplied relation, which is the market supply curve.

The supply curve and its upward slope

Table 3 and Figure 8 illustrate the supply curve. Note that as price increases, quantity supplied also increases; and as price decreases, quantity supplied decreases. Because both quantity supplied and price move in the same direction, the supply schedule has an upward slope.

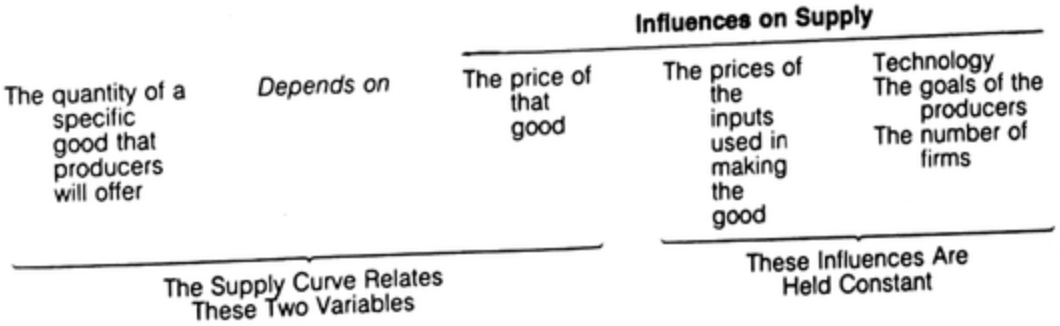


Table 3 A market supply schedule for gasoline

Price	Quantity Supplied (gallons per week)
$3.50	17,400
3.00	15,800
2.50	14,100
2.00	12,300
1.50	10,300
1.00	7,800
.50	5,000

Some of the points that were important in understanding demand are also important for supply. First, the quantity figures must be linked to some specific period of time, such as quantity per week, month, or year. Without this information, the quantity figures would be impossible to interpret. Second, the quantities represented by the supply curve represent the *desired* quantities of sales. They are the amounts producers are willing to offer or sell at each price, not the amounts they actually succeed in selling.

An additional point pertaining specifically to supply should be made. The supply curve—the precise relation between quantity offered and specific prices—exists only for certain types of industries; those in which the producers see themselves as *controlled* by a market price, rather than as *controlling* price in the market. Nonetheless, the general principles of supply—what will increase or decrease each producer's desire to offer various quantities at a certain price—are still useful in discussing supply conditions in all industries.

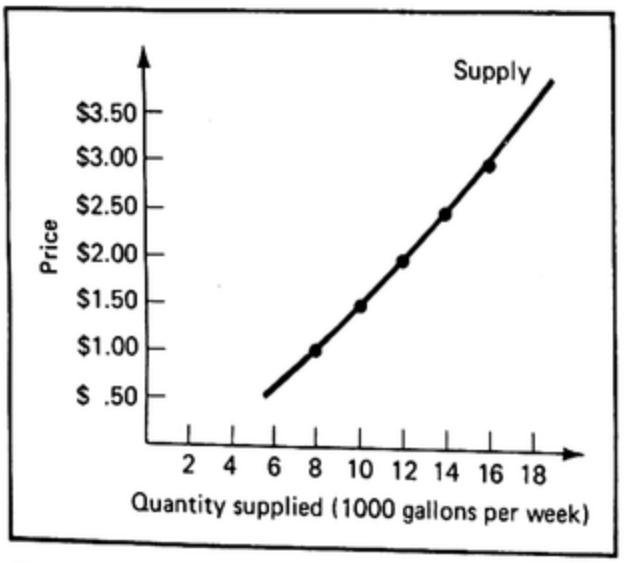


Figure 8 A market supply schedule for gasoline

The supply schedule traces out the quantities that producers are willing to supply at various prices. The upward slope of the supply curve illustrates the fact that higher prices will cause larger amounts of the good to be supplied.

Quantity supplied versus supply

Quantity supplied and *supply* are terms that must be used as precisely as quantity demanded and demand.

Quantity supplied *refers to a specific combination of price and quantity, which is just one point on the supply curve.* A change in quantity supplied is represented by a movement *along the supply curve from one price-quantity combination to another,* such as from *A* to *B* on Supply Curve I in Figure 9. A change in quantity along a supply curve can be caused only by a change in price.

Supply *refers to the entire relationship between prices and quantities supplied; that is, to the entire supply curve. A change in supply is caused by a change in any of the influences on supply, except price.* A change in the cost of an input or in the number of firms, for example, would cause a different quantity to be supplied at any given price. The only way to show this is to shift the entire curve. If the producers will offer more of a good at any given price, then the supply curve will shift to the right. This is shown in Figure 9 by the shift from Supply Curve I to Supply Curve II.

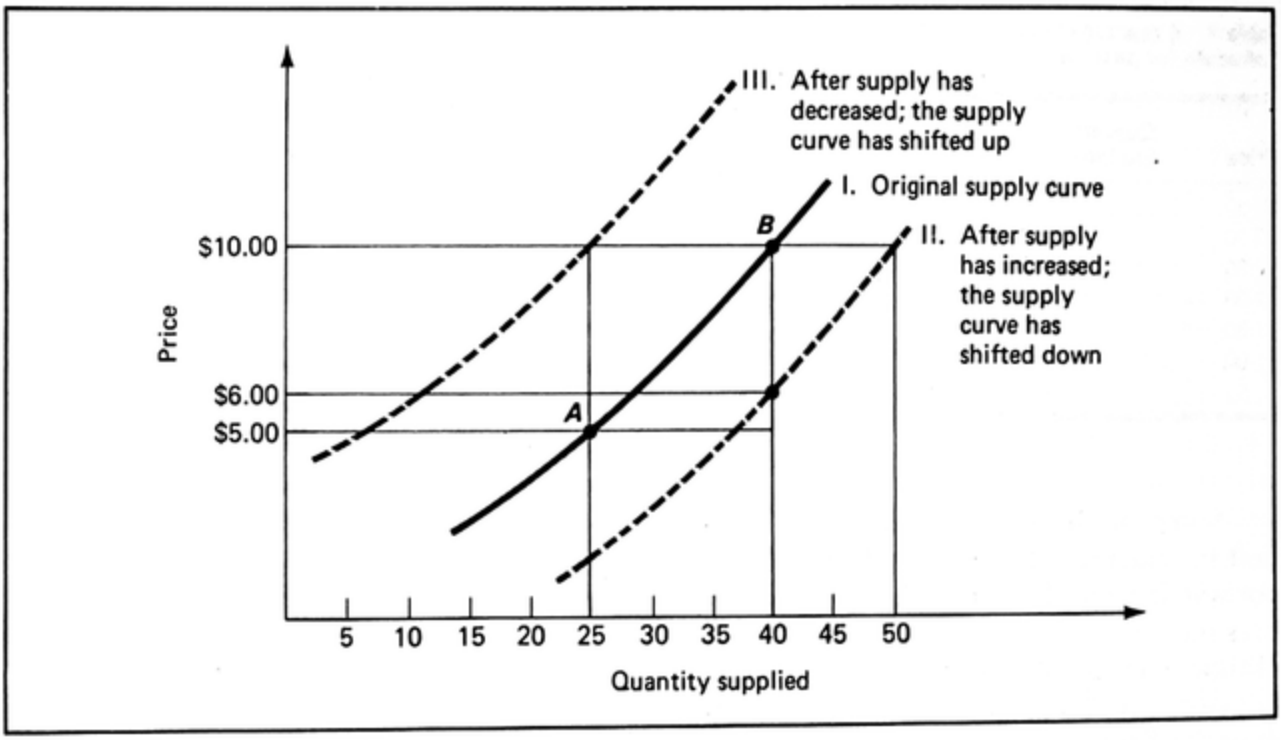


Figure 9 A movement along the Supply Curve differs from a shift of the curve

A change in quantity supplied is caused by a change in price. It is represented by a movement along the supply curve, such as the move from Point *A* to Point *B* on Supply Curve I. A change in supply, caused by a change in any influence on supply except price, means that a different quantity will be supplied at a given price. An increase in supply, with more being offered at a given price, is represented by a rightward shift in the supply curve, such as the shift from Supply Curve I to Supply Curve II. A decrease in supply, with less being offered at a given price, is represented by a leftward shift in the supply curve, such as the shift from Supply Curve I to Supply Curve III.

This shift can be interpreted in two different ways. The increase in supply can be viewed as a larger quantity being supplied at every price. Or, instead, the increase in supply can mean that each quantity is being offered at a lower price. In Figure 9, with supply increasing from Supply Curve I to Supply Curve II, 50 units are now available at a price of $10 instead of 40 units. Or, in the other view: 40 units are now available at a lower price, at $6 rather than $10.

If producers' willingness to supply a good decreases, then the supply curve will shift to the left. Less will be offered at every price; or each quantity will be offered at a higher price. When the curve moves down, or to the right, one says that "supply has increased." If instead the curve shifts up, or to the left, then "supply has decreased." It may seem odd that an *upward* shift in the curve means that supply has *decreased*, but that is correct. To avoid confusion, think in terms of leftward and rightward shifts of the supply curve, rather than upward and downward shifts.

Finally, the supply curve itself can be interpreted in two different ways. Consider Supply Curve I in Figure 9. It shows the highest quantity that will be supplied at each price (40 units at $10), or the lowest price that suppliers will accept for each level of output ($10 for 40 units). Viewed either way, greater levels of output will be offered only at higher levels of price. This is because extra production requires extra effort and cost. How sharp an increase in price is necessary to bring forth a given increase in supply can be measured by the *elasticity of supply*.

Elasticity of supply

The elasticity of supply *measures how responsive the quantity of supply is to changes in price.* The change in price is the cause; the change in quantity the effect. The formula for supply elasticity is:

$$\text{Elasticity of supply} = \frac{\text{percentage change in quantity supplied}}{\text{percentage change in price}} = \frac{\%\Delta Q \text{ supplied}}{\%\Delta P \text{ supplied}}.$$

This elasticity is similar to price elasticity of demand, but there is a difference. Since the supply curve is upward sloping—showing a direct relation between price and quantity—the elasticity of supply will have a positive sign. Both halves of the fraction move the same way. If price increases, quantity supplied will increase. If price decreases, quantity supplied will decrease.

If the elasticity of supply is greater than 1, economists say that the good has an *elastic supply*, with quantity supplied being relatively responsive to changes in price. If the elasticity is less than 1, the good is said to have an *inelastic supply*, with quantity supplied being relatively unresponsive to changes in price. A supply elasticity *equal* to 1 is the borderline case in which the percentage change in price and in quantity supplied are equal.

Elasticity is not identical to slope As with demand, the supply curve's slope is not the same as elasticity. The slope is simply $(P_1 - P_2)/(Q_1 - Q_2)$, while the elasticity, using the midpoint method, would be $\frac{Q_1 - Q_2}{(Q_1 - Q_2)/2} / \frac{P_1 - P_2}{(P_1 - P_2)/2}$. The slope of a straight-line supply curve will be constant, while the elasticity of supply will usually vary from point to point along the curve. That is illustrated in Figure 11 by a straight-line (constant slope) supply curve, whose elasticity values range from 6.0 to 2.0 and 1.6.

But in three special cases, supply curves have a constant elasticity throughout. Two of these cases are shown in Figure 10. At one extreme, shown in Panel I, supply is *perfectly elastic*, represented by a horizontal curve, with elasticity equal to infinity throughout. The other extreme, shown in Panel II, is the case of *perfectly*

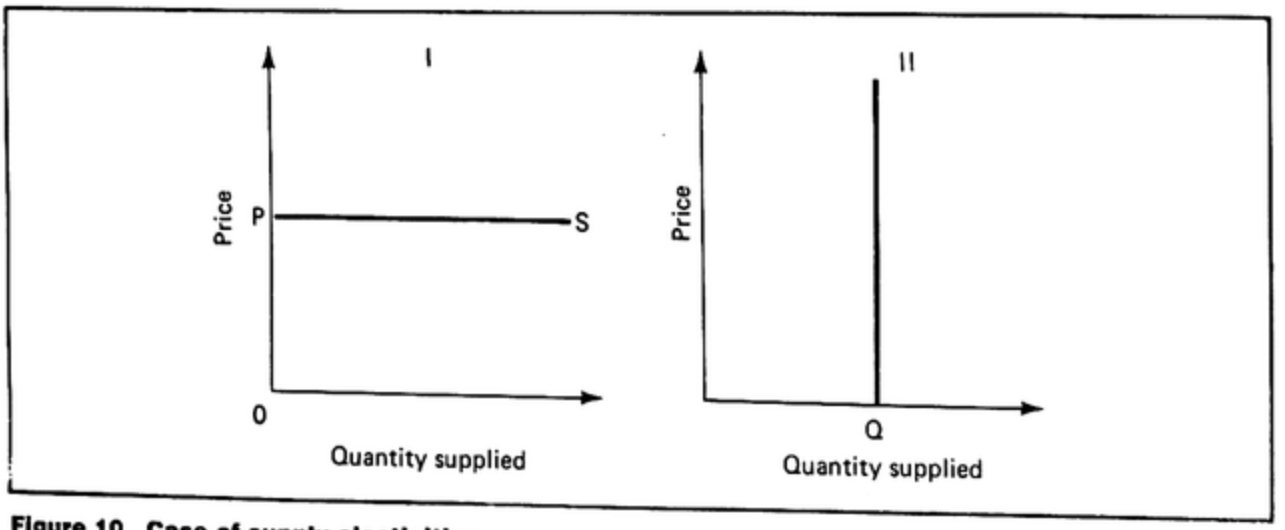


Figure 10 Case of supply elasticities

For an infinitely elastic supply (Panel I), any amount will be supplied at the going price, but nothing will be supplied at a lower price. For a perfectly inelastic supply curve (Panel II), price has no influence on quantity supplied. The supply is fixed at a given level.

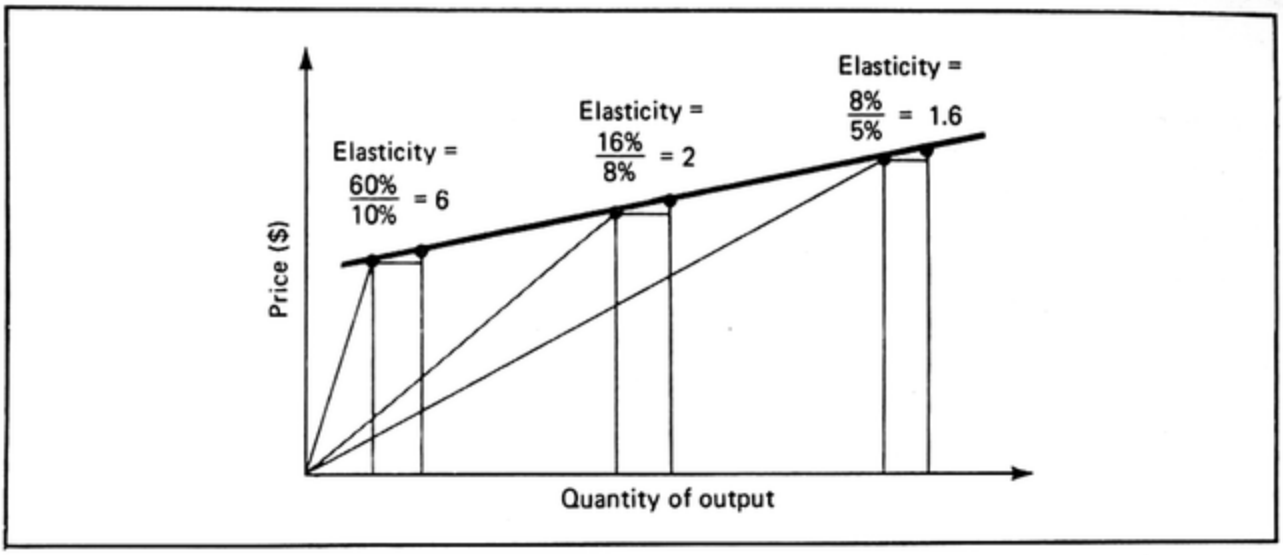


Figure 11 Elasticity varies along a straight supply curve (if it is not a ray or line through the origin)

The variation of elasticity is shown by calculating elasticity at three points. Moving to the right, the elasticity starts high and then declines. Eventually it would go below 1.0. Why does the decline occur? Because the *percentage* rises in price decline more slowly than do the *percentage* rises in quantity.

inelastic supply, in which the supply curve is vertical and has an elasticity of zero throughout. In this case, price has no influence on the quantity supplied. The quantity is fixed, no matter how high or low the price may go. Most actual supply elasticities lie between these two extremes.

The third exception to differing elasticities at each point on the supply curve is the case of a supply curve that lies along a ray from the origin. *For such supply curves, elasticity is constant and equal to 1 at each point.* This is illustrated by Schedules *A* through *C* in Figure 12. Though their *slopes* differ, these schedules all have a constant elasticity equal to 1 throughout their lengths.

Students often find this rule implausible at first, because *A* looks less elastic than *C*. Yet the slopes are constant, and so are the elasticities, along the rays. This constancy contrasts with demand curves, where a straight-line curve has differing elasticities at every point (except for the vertical and horizontal extreme cases).

This unitary value of supply elasticity along rays from the origin is extremely helpful in estimating quickly how elastic a supply curve is. It provides the general rule: Wherever a supply curve is steeper than a ray from the origin passing through it, the curve is *inelastic*. Wherever the curve is flatter than a ray from the origin, it is *elastic*.

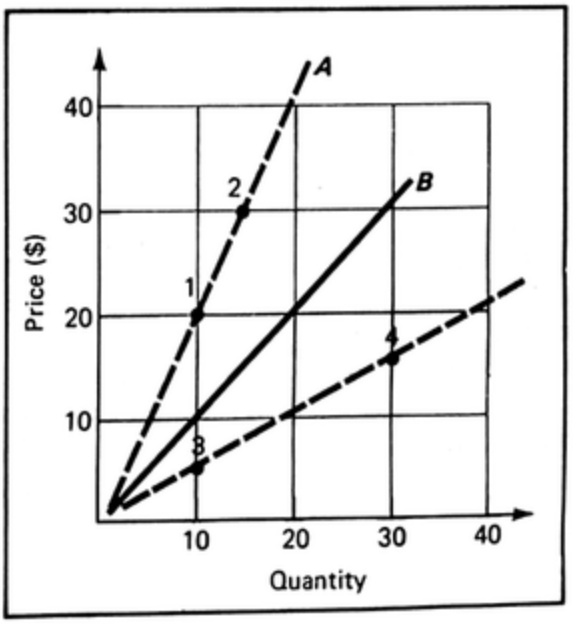


Figure 12 There is constant elasticity of 1 along supply curves that are rays from the origin

This can be verified by calculating the elasticities at points 1 and 2, and at points 3 and 4. The percentage changes in quantity and price will always be identical to each other, so the elasticities will all be equal to 1.

Thus, a quick test for supply elasticity is to draw a ray from the origin through

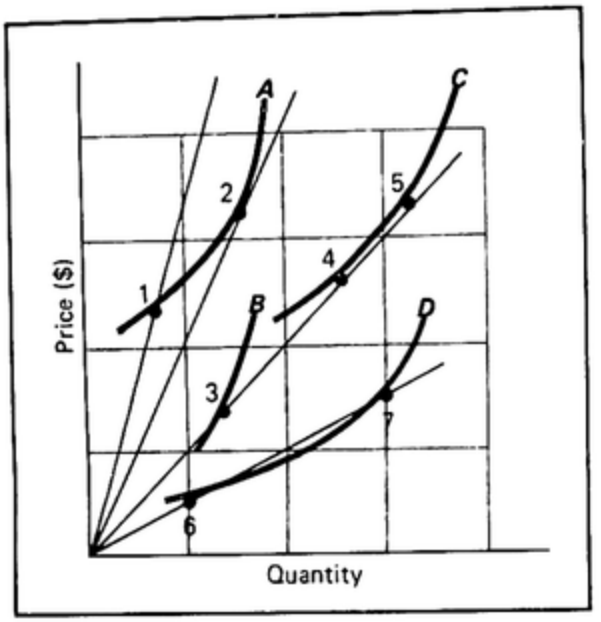


Figure 13 Testing elasticities of supply by using rays from the origin

Consider Curve *A*. The section of the supply curve containing Point 1 is flatter than the ray from the origin drawn through that point. Point 1 must be located on an elastic portion of Supply Curve *A*. Supply Curve *A* is tangent to the ray at Point 2, so its elasticity equals 1 at that point. Along Curve *B*, elasticity is less than 1.0 at point 3. Along Curve *C*, elasticity equals 1.0 in the segment between Points 4 and 5. Along Curve *D*, elasticity is above 1.0 at Point 6 but below 1.0 at Point 7.

whatever part or point of the curve you are interested in. Then just note which is steeper, the ray or the curve. The method also works for *curved* supply schedules, as is shown in Figure 13. If you learn to draw rays lightly on any supply curve diagram, you can quickly show the relative elasticity or inelasticity of any part of any supply curve.

Determinants of elasticity of supply

Four major factors influence elasticity of supply.

1. *Time.* Just as with demand, the elasticity of supply varies directly with the length of the time period: A longer period gives a higher elasticity. The longer time period allows producers to make a fuller adjustment in quantity in response to price changes.

2. *Size of the industry.* Industries that purchase only a small percentage of the total output of their input suppliers can usually buy more inputs quickly with only small increases in price. A large increase in their input use requires only a small increase in total production by suppliers of the input. A local restaurant, for example, can easily expand by hiring a few more waiters and increasing its food purchases, without putting any strain on the entire local labor market or food suppliers. At the opposite extreme is an industry whose demand for inputs accounts for a large percentage of total demand for the input. Growth in the industry, and therefore in its demand for inputs, may severely strain the capacity of the input suppliers. The input suppliers' cost will rise, leading to an increase in the price of inputs.

3. *Special inputs in limited supply.* A specific input that an industry needs may be limited in supply for other reasons. Resource-based industries—such as farm products, lumber, metals, chemicals, or oil—often face this situation. Their supply curves slope up because their key inputs are physically limited. To get more of these inputs, the firms may have to offer steeply higher prices to bid the inputs away from other uses. Or they may have to alter the production process in costly ways to stretch the key inputs further. As a result, the supply curve will be relatively inelastic, showing that more output will be available only at sharply increased prices.

4. *Capital intensity. Heavy industries* use large amounts of capital, such as specialized machinery or equipment, in their production process—for example, blast furnaces in steel plants and oil-refining equipment. Because ordering, receiving, and putting new equipment into operation takes a long time, short-run supply will be relatively inelastic. An increase in price will cause only limited increases in quan-

Supply Elasticity Increases with Time

1. *Coal.* The price of coal rises by 100% because demand increases. Responding to the rise in price, coal companies move their coal stockpiles to market within a week. If this allowed a 10 percent rise in quantity, elasticity for that one-week period would be 10% ÷ 100% = 0.10. In a month, many high-cost coal mines are reopened. Quantity may rise by 30 percent, suggesting an elasticity of 0.30. In a year, 20 entirely new coal mines may have been opened and be starting production. Output is up by 125 percent, suggesting an elasticity of 125% ÷ 100% = 1.25.

2. *Houses.* The price of housing rises by 50 percent because demand shifts upward. Within a week, a few lofts have been spruced up for renting, raising the quantity of new housing per year by 3 percent. Supply elasticity for a one-week period appears to be 3% ÷ 50% = 0.046. In six months, construction has been speeded up to get half-finished houses ready, and the quantity of housing per year is up by 20 percent. Supply elasticity for the period is 0.40. Within a year, the existing stock of housing has risen sharply by 100%, so that the elasticity of supply is 2.0.

tity supplied in all but the longest time periods. In *light* industry, such as small appliances, or local trades, the use of capital is less specialized and extensive, so that capacity is more flexible and can be much more easily adjusted. An increase in price, therefore, can call forth larger increases in quantity supplied. Supply will therefore be relatively elastic.

These four factors—time, size, special inputs, and capital intensity—all work to determine how responsive the quantity supplied will be to changes in price. If supply is relatively elastic, then increases in demand can be accommodated by suppliers with only moderate increases in price. If supply is relatively inelastic, then increases in demand are harder for producers to accommodate. The increase in demand will be met by only moderate increases in quantity and relatively large increases in price.

Interaction of supply and demand

You have now examined both demand (the buyers' side of the market) and supply (the sellers' side of the market). By fitting together demand and supply analysis, you arrive at one of the most important parts of economic analysis: the determination of market price and quantity.

Consider the gasoline market illustrated by the demand and supply information given in Figure 14. Imagine that you have been asked what the equilibrium price and quantity are. You may not even know what "equilibrium prices and quantity" mean, but your attention would probably be drawn to the price and quantity

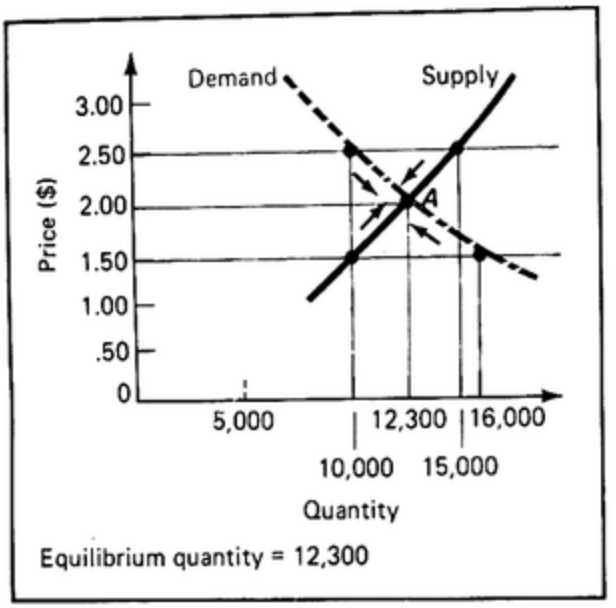


Figure 14 Equilibrium is reached where supply and demand intersect

At a price of $1.50, consumers want to buy far more gallons than suppliers offer. At $2.50, the reverse occurs: Consumers want less than the suppliers offer. Only at Point *A*, at the intersection of the two curves, are the quantities and prices of the two sides—consumers and suppliers—just equal. Note that at any price above equilibrium, there is excess supply. At any price below the equilibrium price, there is excess demand.

Demand		Supply	
Price ($ per gallon)	Quantity (gallons per week)	Price ($ per gallon)	Quantity (gallons per week)
3.50	6,200	3.50	17,400
3.00	8,000	3.00	15,800
2.50	10,000	2.50	14,100
2.00	12,300	2.00	12,300
1.50	15,000	1.50	10,300
1.00	18,000	1.00	7,800
.50	21,500	.50	5,000

combination of $2.00 and 12,300 gallons that exists at the intersection of the supply and demand schedules. That intersection is the only point where quantity supplied equals quantity demanded.

To see why, consider any price above $2.00, such as $2.50. At that price, suppliers will bring 14,100 gallons of gasoline to the market, while consumers wish to buy only 10,000 gallons of it. There is an *excess supply* of 4,100 gallons. What happens? To sell the excess supply, suppliers will have to lower the price of a gallon of gas. As the price of gasoline falls, suppliers offer smaller amounts of gasoline, and the quantity supplied drops. This can be represented by movement down and to the left *along the supply schedule.*

Meanwhile, as the price falls, the consumers want to purchase more gasoline. This increase in quantity demanded occurs as a rightward movement down along the demand curve. The excess supply of gas begins to dwindle. The pressure to reduce the price will continue until the amount of gas that suppliers wish to offer is just matched by the amount of gas that consumers wish to buy. This equality of quantity supplied and quantity demanded will occur only at the point where the two curves intersect, at a price of $2.00 and quantity of 12,300 gallons of gasoline. At that point, the excess supply has been eliminated and there is no further downward pressure on price.

Now suppose that price had originally been below $2.00, such as at $1.50. At that price, consumers wish to purchase 15,000 gallons of gas, while suppliers are willing to offer only 10,300 gallons. There is an excess demand of 4,700 gallons, which shows up as lines of cars at gas stations. Seeing the long lines of unsatisfied customers, the station operators will quickly realize that there is a greater demand for gasoline than they can meet. In a free market, they will respond to this shortage by increasing price to try and raise their profits. As the price of a gallon of gasoline increases, suppliers will be led to offer more gasoline, moving to the right along their supply curve. Consumers will then respond to the higher price by reducing the amount of gas they are willing to purchase, moving to the left along their demand curve.

The price will continue to rise, along with the increase in quantity supplied and the decrease in quantity demanded, until

the point is reached at which suppliers are willing to offer exactly the amount that consumers are willing to buy. Once again, the price and quantity gravitate to the combination represented by the intersection of supply and demand—a price of $2.00 and a quantity of 12,300 gallons. The price and quantity will now remain stable until some further change in supply or demand conditions occurs.

The intersection of the supply and demand curve represents market equilibrium. ***Equilibrium is a balancing of forces. At the point where the supply and demand curves intersect, supply and demand forces are perfectly balanced, so that the quantity supplied equals the quantity demanded.*** If the market is operating at some point other than the equilibrium, the excess supply or excess demand sets in motion the market forces that will cause price and quantity to move back toward equilibrium. Once that equilibrium is reached, the price and quantity will tend to persist until there is a change in some underlying supply or demand condition. This would be represented by a *shift* in the supply or demand curve. The market price and quantity would then move toward the new equilibrium.

Such a shift can easily be illustrated, as in Figure 15. Suppose that demand for gasoline has decreased because the price of automobiles (a complementary good) has risen sharply. (Note: The curve need *not* shift parallel to its original position.) At the old equilibrium price of $2.00, consumers wish to buy only 8,000 gallons, though suppliers are still ordering 12,300 gallons for sale. Excess supply is 4,300 gallons. To dispose of it, the suppliers cut the price and reduce their orders. Moving to the left down the supply curve, they finally reach Point *B*, the new equilibrium. Price is $1.40 and quantity is 10,000 gallons. (Incidentally, the elasticity of supply over this part of the supply curve is $[-20.6\%] \div [-35.3\%] = 0.58$.)

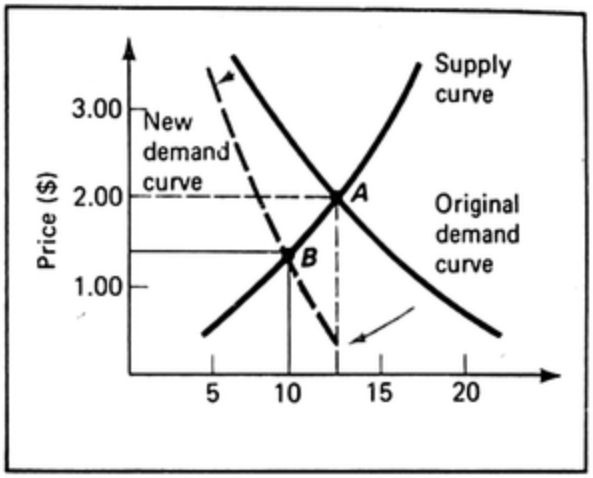


Figure 15 A shift in demand causes a new equilibrium

The original equilibrium is at Point *A*. Then demand shifts down. At a price of $2.00, there is a physical surplus of 4,300 gallons, because buyers want only 8,000 gallons, but suppliers are offering 12,300 gallons as before. The new equilibrium is reached at Point *B*, where both price ($1.40) and quantity (10,000 gallons) are lower.

A shift in supply causes a similar adjustment with a movement along the demand curve to the equilibrium. The new equilibrium will always be the new intersection of the demand and supply curves.

These steps lead to an outcome that fits economic logic. As Figure 15 shows, the fall in demand has caused both price and output to decrease. Consumers' desire for gasoline is now less than before, so that the fall in price and quantity bought is precisely what you would expect. Moreover, the change has come about spontaneously, by market actions that cause the new equilibrium to be reached.

One last point: There is nothing morally "good" about equilibrium. Price and quantity are not necessarily fair or just or equitable in any normative sense. Market equilibrium is simply the price and quantity combination at which the market clears, with the quantity supplied equal to the quantity demanded.

Summary

This chapter examines the demand and supply sides of the market separately and then brings them together to show how the market equilibrium is determined. The main points in the chapter are:

1. Supply and demand forces interact in markets. Each market is defined by both the nature of the product and the geographic area in which the product is sold.
2. The factors that influence demand include: (a) price; (b) income; (c) preferences; (d) price of other goods, such as substitutes and complements; (e) population; (f) income distribution; and (g) expectations about future prices.
3. If all influences on demand are held constant, and only price is allowed to vary, then the relation between price and quantity demanded can be isolated and studied. This relation is shown by the *demand curve.*
4. *Quantity demanded* refers to a particular point on the demand curve that represents a specific price-quantity combination. A *change in quantity demanded,* caused by a change in price, refers to a movement along the demand curve from one price-quantity combination to another. *Demand* refers to the entire price-quantity relationship; the entire demand curve. A *change in demand* refers to a shift in the entire curve, caused by a change in any influence on demand except price.
5. *Price elasticity of demand* $\left(\frac{\%\Delta Q}{\%\Delta P}\right)$ measures the relative responsiveness of quantity demanded to changes in price. Because of the inverse relation between changes in price and quantity, price elasticity will always be negative.
6. There is an important relation between price elasticity and total revenue *(TR = P × Q).* If demand is elastic, total revenue will move in the direction of quantity changes. If demand is inelastic, total revenue will move in the direction of price changes. If price elasticity is equal to 1, a change in price and quantity causes no change in total revenue.
7. The main influences on price elasticity are: necessity, the number of available substitutes, time, and the percentage of income spent on the particular good.
8. *Income elasticity* $\left(\frac{\%\Delta Q}{\%\Delta \text{Income}}\right)$ measures the responsiveness of quantity demanded to changes in income. A *positive* income elasticity indicates that the good is a *normal* good, with quantity demanded increasing as income increases. A *negative* income elasticity indicates that the good is an *inferior* good, with quantity demanded decreasing as income increases. If the value of income elasticity is greater than 1, the good is said to be a normal good with an *income-elastic demand:* Quantity demanded is relatively responsive to changes in income. If the elasticity value is between 1 and zero, the good is said to be a normal good with an income-inelastic demand: Quantity demanded is relatively unresponsive to changes in income. A value less than zero indicates an inferior good.
9. Cross-elasticity of demand $\left(\frac{\%\Delta Q \text{ of Good 1}}{\%\Delta P \text{ of Good 2}}\right)$ measures how the

quantity demanded of one good responds to price changes of another good. A *positive* cross-elasticity indicates that the goods are *substitutes*. A *negative* cross-elasticity indicates that the goods are *complements*. Complementary goods are used together, as, for example, cars and gasoline.

10. Influences on supply include: (a) price; (b) costs, which depend on costs of inputs and technology; (c) goals of firms; (d) number of firms; (e) changes in the price of goods related on the production side; and (f) producers' expectations about future prices.

11. If all influences on supply are held constant and only price is allowed to vary, the quantity supplied–price relation can be isolated and examined. This relation is illustrated by the *supply curve*. The curve has a positive slope, showing that more of the good will be supplied at higher prices.

12. *Quantity supplied* refers to a particular point on the supply curve that represents a specific combination of price and quantity. A *change in quantity supplied* refers to a movement along the supply curve from one price-quantity combination to another, caused by a change in price. *Supply* refers to the entire supply curve relationship. A *change in supply* refers to a *shift* in the supply curve, caused by a change in any influence on demand except price.

13. Elasticity of supply $\left(\frac{\%\Delta Q \text{ supplied}}{\%\Delta P}\right)$ measures the relative responsiveness of quantity supplied to changes in price. Because price and quantity supplied are directly related, supply elasticity will be positive. If elasticity is greater than 1, supply is said to be elastic, with quantity supplied being relatively responsive to changes in price. In most cases, supply elasticity will differ from one point on the supply curve to another.

14. The important influences on supply elasticity are: time, size of the industry, flexibility of supply of inputs, and capital intensity. A summary of the main types and ranges of both demand and supply elasticity is provided in Table 4.

Table 4 The main types and ranges of elasticities

	Phrase	Value of the Elasticity Ratio
Demand		
1. *Elasticity of Demand*	Elastic	Greater than 1
	Unitary elastic	Equal to 1
	Inelastic	Less than 1
2. *Income Elasticity*	Normal Good	Greater than zero (positive)
	Inferior Good	Less than zero (negative)
	Income elastic	Greater than 1
	Income inelastic	Less than 1
3. *Cross-elasticity of Demand*	Substitutes	Greater than zero (positive)
	Unrelated	Zero
	Complements	Less than zero (negative)
Supply		
1. *Elasticity of Supply*	Elastic	Greater than 1
	Unit elastic	Equal to 1
	Inelastic	Less than 1

15. The market price and quantity are determined by the interaction of demand and supply. The price-quantity combination at the intersection of the supply and demand schedules is the *equilibrium* price and quantity. This point of market equilibrium is the only one at which the quantity supplied and the quantity demanded are equal.

Key concepts

Market
Substitution effect
Income effect
Demand
Quantity demanded
Price elasticity of demand
  elastic demand
  inelastic demand
  unitary demand
Income elasticity of demand
  normal good
  inferior good
Cross elasticity of demand
  substitutes
  complements
Quantity supplied
Supply
Elasticity of supply
Market equilibrium

Questions for review

1. Consider the following list of goods and services. Would you place them in a local or national market? Explain your reasoning.

bread	milk
automobiles	bicycles
fresh tomatoes	doctors' services
small appliances	cameras

2. Suppose that you are trying to determine what influences the quantity of jogging shoes purchased in the United States. List the factors that you feel would be important influences on the demand for jogging shoes. Since you are working with a very specific good, you should be able to add specific influences to the list of general influences given in this chapter.

3. List three goods for which you feel your demand is elastic, and three for which your demand is relatively inelastic. By comparing your lists, can you determine the factors that influenced your elasticity of demand for these goods?

4. Suppose that the trustees of your college decide that the college needs more revenue. They therefore vote for a 15 percent hike in tuition. Will the tuition hike necessarily accomplish their objective? Explain.

5. List three goods that you would classify as inferior goods and three goods that you consider to be normal goods. If you were earning $50,000 a year instead of living on a student income, would you still consider all of these last three goods to be normal goods? Explain.

6. Thinking of your own purchasing patterns, name some goods that you consider substitutes, and some goods you consider complements.

7. Suppose that you were grading exams and came across the following statement: "Quantity supplied refers to the amount that producers sell to consumers at various prices." Could you give full credit for the explanation? Why or why not?

8. Would you expect supply to be relatively elastic or inelastic in the following cases?

a. steel industry
b. lemonade stand
c. wheat farming
d. canned fruit
e. local doctors' services
f. local real estate services

9. Each of the following examples represents a change that will affect the wheat market. Using supply and demand diagrams, explain how each change is likely to affect the supply and demand side of the wheat market. Make sure that you indicate how equilibrium price and quantity are likely to be affected, and explain carefully whether the change will be either a change in quantity supplied or demanded, or a change in supply or demand.
   a. There is an exceptionally large harvest of wheat.
   b. Consumers hear that some wheat supplies have been contaminated with a toxic substance.
   c. The cost of fertilizer used by wheat farmers increases.
   d. Consumers develop an increased desire for products made with wheat.
   e. Excess supplies of wheat are building up.
   f. An increased desire for rural living causes more people to become wheat producers.
   g. The population of the United States increases.
   h. Substantial improvements in the technology of wheat farming occur.
   i. U.S. income per capita increases steadily.
   j. The price of soybeans increases while the price of wheat remains constant. (Farmland can be used for either wheat or soybeans.)

5

Demand and Supply in Action

As you read and study this chapter, you will learn:

- the simple analysis of demand and supply in equilibrium
- how lessons of demand and supply work in practice
- how to analyze the effects of distorting the market solution, in four practical cases
- the simple problems of measuring demand and supply in real markets

You have probably discussed subjects about which people hold strong views, such as U.S. foreign policy, the welfare system, or minimum wage legislation. What starts as a friendly conversation can become an angry debate if people begin to exchange value judgments and opinions rather than facts.

To call the minimum wage law a harmful interference with free markets is a provocative statement of opinion that invites a heated response. To say that the minimum wage can contribute to inflation is a little more specific and invites a more reasoned response. To say that the minimum wage can raise prices and reduce the number of jobs, and then to explain why that might be so, is even more positive and less normative. By using economic analysis to reach specific answers, one has the best chance of encouraging an equally reasoned and rational response. One is discussing specific theories and facts rather than airing broad conclusions.

People often know more theory or facts about economic top-

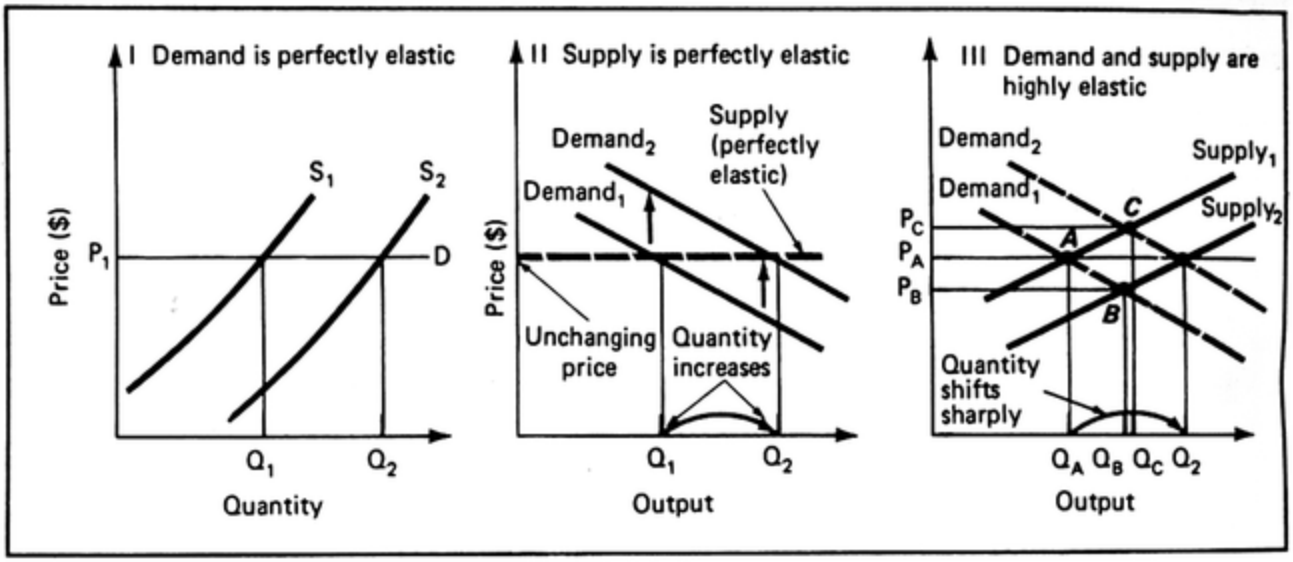


Figure 1 Elastic demand and supply

In Panel I, perfectly elastic demand is represented by a horizontal line. An indefinitely large amount will be demanded at the present market price, while nothing will be demanded at a higher price. A change in supply, such as the shift from S_1 to S_2, will cause a change in quantity (from Q_1 to Q_2) but no changes in price. In Panel II, perfectly elastic supply is represented by a horizontal line. An indefinitely large amount will be supplied at the present market price. A change in demand, such as the shift from D_1 to D_2, will cause a change in quantity (from Q_1 to Q_2) but no change in price. P_A and Q_A represent the equilibrium price and quantity when market conditions are represented by D_1 and S_1. Think of possible changes in the market. Demand might stay at D_1, while supply increased to S_2. Market equilibrium would move from A to B. Price would move from P_A to P_B, while quantity would move from Q_A to Q_B. Or supply might stay at S_1, while demand increased from D_1 to D_2. Market equilibrium would move from A to C. Price would move from P_A to P_C, while quantity would move from Q_A to Q_C. In both cases, note the relatively large changes in quantity.

ics than they realize. But it takes practice to learn how to apply theory to such issues and to use the most relevant information. This chapter is devoted to just that kind of practice. It applies the supply and demand theory introduced in Chapter 4 to various issues, such as the impact on market price and quantity of supply and demand changes, the effects of minimum wage legislation, and the relative burdens of sales taxes. When you have finished the chapter, you should have a better feeling both for the tools of supply and demand analysis and for how to apply theory to a specific problem.

Effects of elasticities on market outcomes

Supply and demand analysis helps determine the impact of market changes on both price and quantity. These market changes show up as shifts in supply and demand schedules. The size of the resulting price and quantity changes depends not only on the magnitude of the shifts, but also on the elasticity or shape of the supply and demand schedules. The analysis of many important issues, such as the impact of the minimum wage on workers' income or the share of a sales tax borne by consumers, depends crucially on elasticity.

You know from Chapter 4 that both supply and demand elasticities can vary from zero to infinity. The market adjustment process is the same, regardless of the elasticities, but the sharpness of the market changes will depend closely on those elasticities. The classic way to grasp the role of elasticities is to consider first the four extreme cases of elasticity, as follows.

Elastic demand and supply

If the demand curve is elastic, quantity changes relatively more than price, in per-

centage terms. Panel I of Figure 1 illustrates perfectly elastic demand. Any amount of the good up to infinity will be bought at P_1 but nothing will be bought at a higher price. Here, market price is determined by demand, and quantity is determined by the location of the supply schedule.

As you will see in Chapter 9, certain firms face a perfectly elastic demand schedule. They can sell as much as they wish at the going market price. While demand is seldom perfectly elastic, the more elastic the demand for the good, the bigger the change in quantity and the smaller the change in price that will result from a given change in supply.

For *perfectly elastic supply*, as shown in Panel II of Figure 1, *the price is determined by supply, and the quantity is determined by demand.* Changes in demand will cause quantity changes but not price changes. The more elastic supply is, then, the larger the changes in quantity and the smaller the changes in price that will result from a given change in demand. Easy adjustments in output resulting in very elastic supply are typical of many narrowly defined, small-volume products, such as candies, small metal products, dresses, and printing. For such goods, output can be increased quickly, simply, and inexpensively in response to changes in demand.

If supply and demand are *both* highly elastic, as shown in Panel III of Figure 1, shifts in supply and demand will cause relatively large fluctuations in quantity and relatively small changes in price. The large changes in quantity occur because consumers can easily adjust their rate of consumption to price changes, and suppliers also have no difficulty adjusting the quantity they supply to changes in demand. Industries with relatively elastic demand and supply include small appliance manufacturers, restaurants, grocery stores, and other retail trades.

Inelastic demand and/or supply

If demand is inelastic, price changes relatively more than quantity, in percentage terms. In a *perfectly inelastic demand schedule*, as in Panel I of Figure 2, consumers will buy the same amount of the good regardless of price. Market price is determined entirely by supply conditions, but quantity is determined by demand. While perfectly inelastic demand is rare, the more inelastic the demand, the larger the change in price that will result from a given change in supply. Extremely inelastic demand would be found for goods such as medicines or textbooks for which there are no practical substitutes.

A *perfectly inelastic supply schedule* is shown in Panel II of Figure 2. Since quantity is fixed, a change in demand can only affect price, not quantity. *Price is determined by demand, while quantity is determined by supply conditions.*

Goods with an inelastic supply, whose quantities are fixed, include land in specific locations, Old Master paintings, antique furniture, and vintage wine. A Rembrandt that sold for $200,000 in 1950 cost $7 million in 1982 because demand moved up a nearly vertical supply curve. Yet, even when supply *seems* to be completely inelastic, the increases in price caused by increases in demand may induce some increase in supply. Swamps, for example, can be drained to create more arable land. The more inelastic the supply, however, the greater the change in price and the smaller the change in quantity that will result from a change in demand.

When demand and supply are *both* highly inelastic, as in Panel III of Figure 2, shifts in supply and demand will cause relatively large changes in price and small changes in quantity. For example, both supply and demand are highly inelastic in the market for U.S. agricultural products. The demand for many foods is fairly inelastic. Even if the price of bread, milk, or

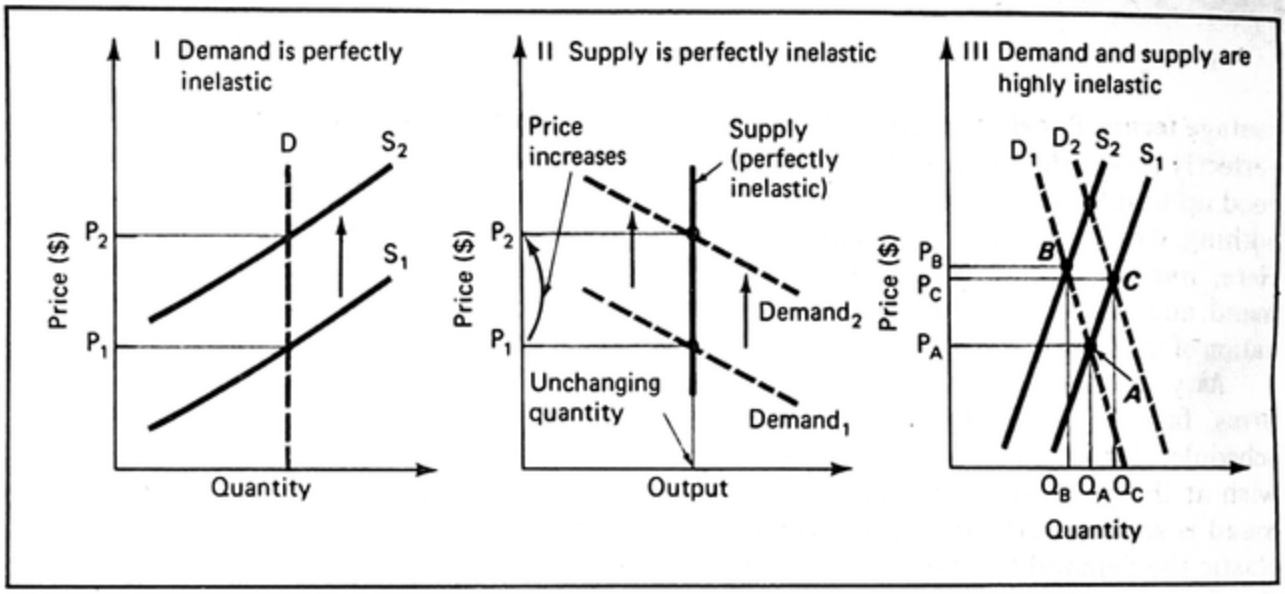


Figure 2 Inelastic demand and supply

In Panel I, perfectly inelastic demand is represented by a vertical line. The same quantity will be purchased regardless of price. In such cases, a change in supply, such as the shift from S_1 to S_2, will cause a change in price (from P_1 to P_2), but no change in quantity demanded. In Panel II, perfectly inelastic supply is represented by a vertical line. The same quantity will be supplied regardless of price. In such cases, a change in demand, such as the shift from D_1 to D_2, will cause a change in price (from P_1 to P_2), but no change in quantity supplied. In Panel III P_A and Q_A represent the equilibrium price and quantity when market conditions are represented by D_1 and S_1. Think of possible changes in the market. Demand might stay at D_1, while supply decreased from S_1 to S_2. Market equilibrium would move from Point *A* to Point *B*. Or, supply might stay at S_1 while demand increased from D_1 to D_2 Market equilibrium would move from Point *A* to Point *C*. In both cases, note the relatively large change in price and relatively small change in quantity that occur when supply or demand shifts.

potatoes is cut in half, there is a limit to how much of these foods people want to eat. Therefore, changes in price will call forth relatively small changes in the quantity demanded.

The supply of agricultural products is also inelastic because significant adjustments in crops and livestock must wait for a new planting or breeding season. Even longrun adjustments are limited because farmland cannot easily be expanded.

The combination of inelastic supply and demand can cause a paradox: Good harvests can result in low farm incomes, poor harvests in high ones. You might find this baffling without a knowledge of elasticity. But if you understand that elasticity measures the responsiveness of quantity to price changes, the paradox is easily resolved. (Try out the paradox on a friend who has never studied economics. You will see how much you have already learned.)

Suppose that S_1 and D_1 in Figure 3 represent normal or average supply and demand conditions in the market for wheat. (Ignore government price supports and assume that market forces operate freely.) Now suppose that there is an especially plentiful harvest. The supply schedule for wheat would shift to the right, as shown by the shift from S_1 to S_2 in Figure 3. To get people to buy the increased amount of wheat, the market price of wheat will fall. The effect of this quantity increase and price decrease on total revenue depends, as you know, on the elasticity of demand. If demand is relatively inelastic or less than 1, as it is for agricultural products, the percentage change in price will be greater than the percentage change in quantity: Farm incomes will fall. When crops are poor, as represented by S_3, market price will increase in response to this decrease in quantity, as shown by P_3 and Q_3. Since the percentage change in price dominates, total revenue will increase. You can see the influence of the size of the

total wheat supply on farmers' income by comparing the rectangles representing total revenue for both a good harvest (P_2 B Q_2 0) and a poor one (P_3 F Q_3 0). The rectangle representing total revenue for the relatively poor harvest (P_3FQ_30) is clearly larger.

The fall in income when crops are good and the rise in income when crops are poor are intensified by the inelasticity of supply. Even with the inelastic demand, shifts in supply would have a milder impact on both price and quantity if supply were more elastic. This is illustrated in Panel II of Figure 3. A set of elastic supply curves (S_E and S_E') and a set of inelastic supply curves (S_I and S_I') are drawn for equivalent changes in quantity. Note that the same horizontal shift in supply from A to B results in a much smaller change in price and quantity when supply is elastic than when it is inelastic. For the more elastic supply, the price change will equal $P_A - P_E$ and the quantity change will equal $Q_A - Q_E$. For the inelastic supply, the price change is $P_A - P_I$ and the quantity change is $Q_A - Q_I$.

These examples, based on differing elasticities, show how important elasticities are in determining the outcome of actual cases. By using the concept of elasticity carefully and frequently, you can develop good judgment about the likely outcomes of market changes in many familiar markets.

The rest of this chapter gives examples of other situations in which supply and demand analysis is important. In some cases, the elasticities and other market conditions are so well known that the analysis can give precise answers. In others we have to estimate market conditions or trends, because the data are less complete. The results can only be approximate, such

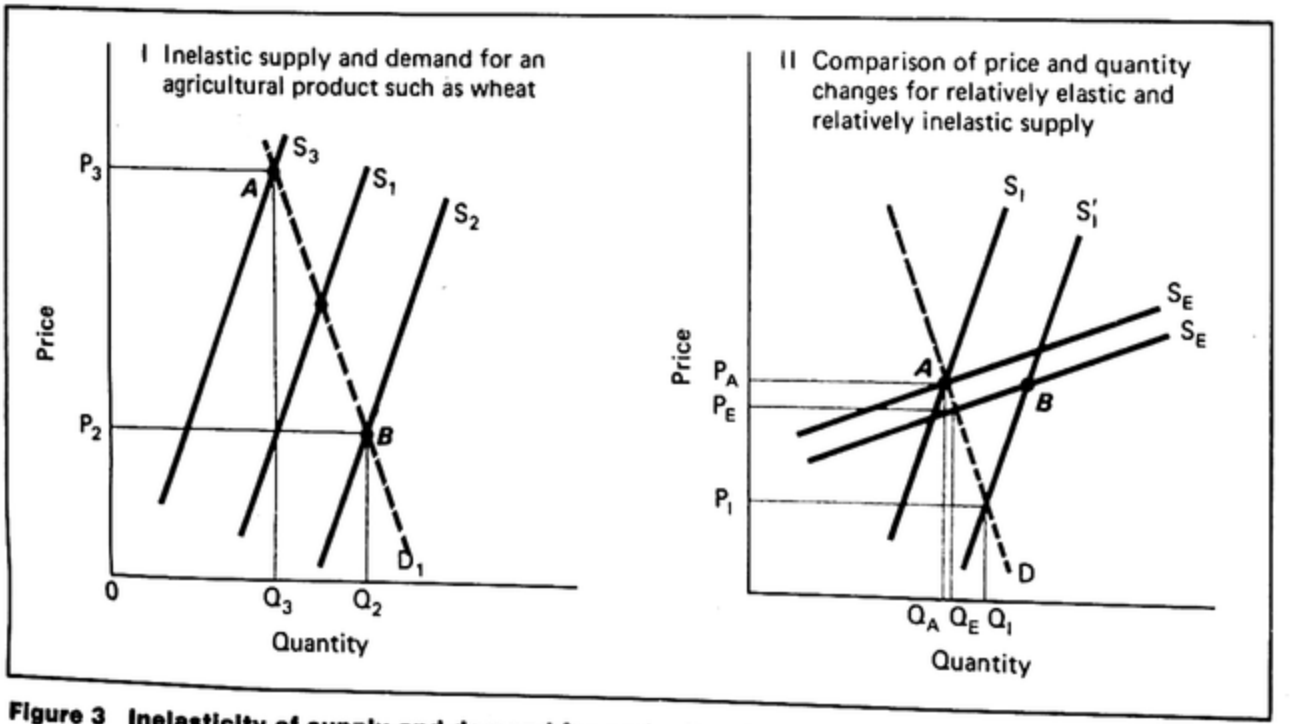


Figure 3 Inelasticity of supply and demand for agricultural products

S_E represents a relatively elastic supply schedule, while S_I represents a relatively inelastic supply schedule. When both are shifted by the same horizontal distance representing a change in quantity from A to B, note the different impacts on equilibrium price and quantity. The shift of the elastic schedule causes a relatively small change in price (P_A to P_E) and quantity (Q_A to Q_E). The shift of the inelastic schedule causes a much larger change in price (P_A to P_I) and quantity (Q_A to Q_I).

as: "If supply is elastic and stable, then oil prices will rise slowly." Yet, even such inexact predictions are often valuable.

The examples that follow suggest the great variety of problems that simple economic analysis can clarify. Some of your own personal choices in markets may be sharpened by using supply and demand analysis. On a larger scale, the same type of analysis can also clarify even the most urgent national and global issues. The first example (in the second main section) shows how a market works when it is left to function on its own. The second set of examples (in the third main section) illustrates how outside interferences in the market process can affect the market outcomes involving price and quantity.

The price of oil

Since 1972, the price of oil has risen sharply, to $32 per barrel in 1980 from less than $3 per barrel in 1972. OPEC, the oil producers' organization, may have influenced that rise, but since 1980, OPEC has had little effect.*

Let us assume that oil supply will be essentially competitive in the future, so that oil prices will now depend solely on shifts in the supply of and demand for oil. Present demand and supply are illustrated by curves D_1 and S_1 in Figure 4. No one can predict with certainty how these might shift. But economic analysis can contrast alternative sets of circumstances, to illustrate the range of likely outcomes for a given year, such as 1990.

First consider a "low price" case. Suppose that conservation efforts by oil consumers are thorough and effective, and that other energy sources such as solar heat and fusion power are developed.

*In Chapter 21 we consider whether the 1970s oil price rise occurred for basic reasons, rather than because of OPEC's influence.

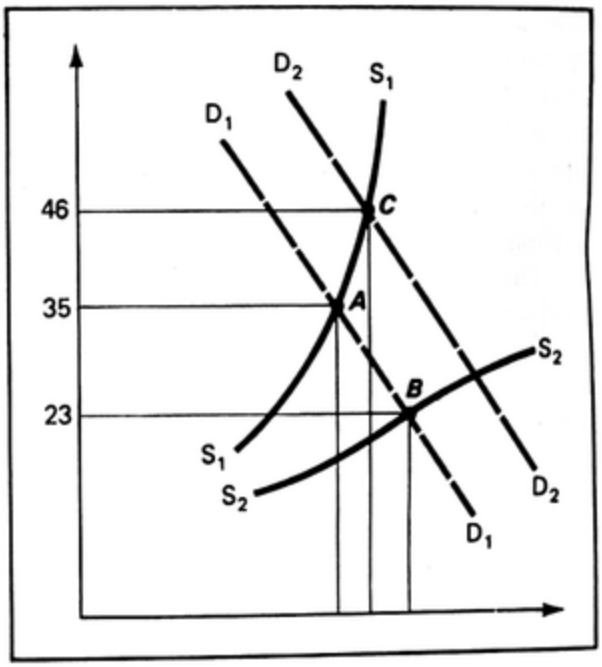


Figure 4 Shifts and elasticities govern the future price of oil

Point *A* is the present market equilibrium for oil. Point *B* is the year 2006 market equilibrium, which would result from a combination of shifts in supply and demand conditions favorable for the United States. Supply increases and becomes more elastic, while demand grows slowly compared with recent decades. Point *C* is the year 2000 market equilibrium, which results from an unfavorable set of supply and demand changes. Supply stays inelastic and grows little, while demand continues increasing at its present rate. Point *C*'s price is twice as high as the price at Point *B*.

These events would slow down the rate of growth in the demand for oil. Suppose, for simplicity, that demand in the year 2000 is identical to D_1 in Figure 4, the present curve. On the supply side, suppose that the search for more oil and gas is highly successful. This increase in supply can be represented by a rightward shift in the supply curve from S_1 to the blue line S_2.

The market equilibrium would move from Point *A* in 1980 to Point *B* in the year 2000, where the new demand and supply curves D_2 and S_2 intersect. The new equilibrium at Point *B* shows more oil available at a lower equilibrium price than at the original equilibrium. Yet this optimistic forecast holds only if there are favor-

able conditions on both the supply and demand sides of the market.

Now consider a "high price" result. Suppose that further conservation efforts are ineffective. Moreover, the development of other fuel sources is slow. Given these circumstances, the demand for oil may rise 5 percent per year. In Figure 4, Demand Curve D_2 illustrates the demand for oil in the year 2000. Meanwhile, on the supply side, suppose that the discoveries of new oil fields are sparse. Supply in the year 2000 would not be much greater than in 1980—about the same as at present, represented by Supply Curve S_1. The market equilibrium would move from Point A to Point C in the year 2000, where D_2 and S_1 intersect. The year 2000 market equilibrium for oil shows less oil supplied than at present for a much higher price per barrel.

Interferences with the market process

So far, we have dealt with cases in which the market was allowed to function freely. The following examples deal with interference with the market process and show how the free market result can be modified or displaced.

There are three main kinds of interferences with the market process: taxes, price controls, and quantity controls. Because they alter the results given by voluntary exchange, these actions may cause important distortions.

The burden of a sales tax

Sales taxes (often called excise taxes) are imposed in most states of the United States. While they raise revenue, they also influence the market process. Who really pays this tax? You pay at the cash register, but the producer of the good might also bear some or all of its burden. With supply and demand analysis, it is easy to see how sales taxes affect both consumers and producers. Elasticity allows you to show who will bear the biggest portion of the tax.

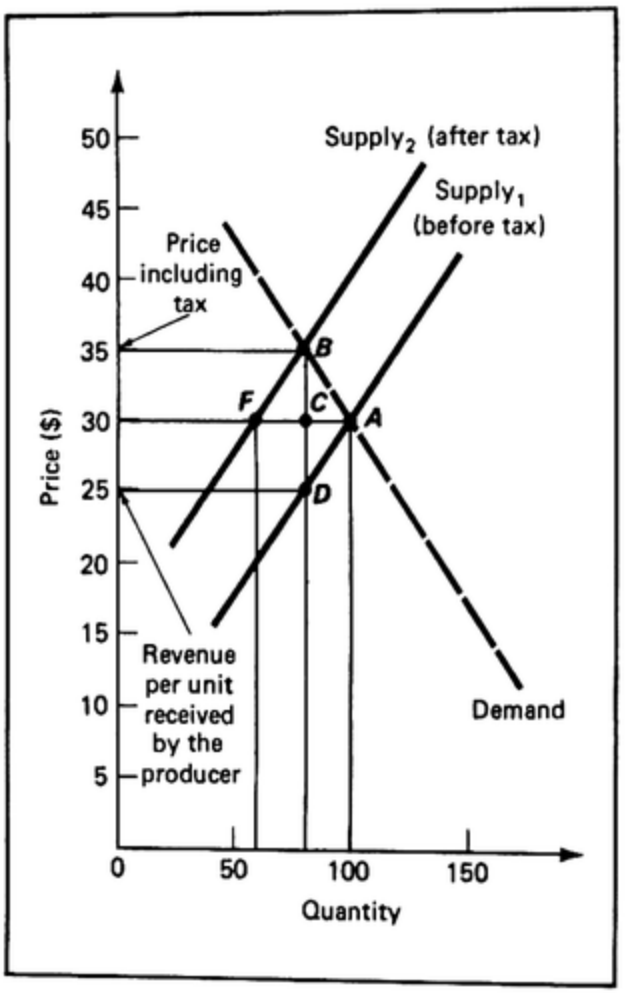


Figure 5 Market equilibrium before and after the imposition of a sales tax.

Consider Figure 5. The free market equilibrium price and quantity are $30 and 100 units. This before-tax equilibrium is represented by Point A. Now suppose that a tax of $10 per unit is imposed on this good. What happens? Since the producer will need to receive the same price for any given quantity supplied, with or without the tax, *the supply curve simply shifts up at every point by $10.* In short, the tax is simply added to the costs of production.

One hundred units will now be supplied to consumers at a price of $40: $30 for the producer (the same revenue for each unit as before the tax) and $10 for the government. This $10 increase in the supply price for each unit is represented by the leftward shift of the supply curve from $Supply_1$ to $Supply_2$.

Now what happens? At the old price of $30, quantity demanded at Point *A* is greater than the amount that producers are willing to supply (Point *F*). This shortage of the good at a price of $30 causes the price to be bid up. The price rises until the quantity supplied is once again equal to the quantity demanded. This new equilibrium will occur at Point *B*, reflecting a higher price ($35) and a lower quantity (80 units) than before the tax. Although the tax was $10, the price per unit rose only $5 (from $30 to $35). That means that consumers are not bearing the full burden of the tax. Some of that burden must also fall on the producers.

To see precisely what the consumers' and producers' shares of the tax are, look again at the new equilibrium at Point *B* in Figure 5. The vertical distance between the old and new supply curve at the new equilibrium quantity of 80 units, measured by *B*–*D* on the diagram, represents the $10 tax. The consumers' portion of the tax is simply the difference between the old and new market price, represented by *B*–*C* on the diagram. In this example, *B*–*C* must be equal to $35 – $30 = $5. Now if the entire tax is $10, represented by *B*–*D*, and the consumers' portion is $5, represented by *B*–*C*, then the producers' share of the tax must be the remaining segment of *C*–*D*. This would be equal to the $10 tax minus the $5 consumers' share, or $5. The *C*–*D* segment is simply the difference between the pretax producer revenue of $30 and the posttax producer revenue of $25. The decrease in revenue per unit represents the producer's share of the tax.

But on whom will the tax burden fall most heavily? The answer depends on the elasticity of demand and supply.

Figure 6 depicts various cases of demand elasticity. Panel I illustrates perfectly inelastic demand. Here, the price increase, representing the consumers' burden of the tax, is equal to the entire vertical distance between the supply curves—in other words to the full amount of the tax. Panel II represents perfectly elastic demand. Here, the upward shift of the supply curve resulting from the imposition of a tax causes no change in price at all. Consumers bear none of the burden of the tax.

Most cases of demand elasticity lie between these two extremes. If you keep in mind that consumers pay all of the tax if demand is perfectly inelastic and none of the tax if demand is perfectly elastic, then it should be easy to see how the consumers' burden of the tax varies with demand elasticity. The more inelastic the demand, the greater the tax burden borne by the consumer, other things being equal. This should make intuitive sense to you, since inelastic demand implies that consumers would rather pay higher prices than reduce their consumption of the good by very much. On the other hand, the more elastic the demand for a good, the smaller the burden of the sales tax that consumers will bear, other things equal. In this case, consumers will more readily give up a good rather than pay higher prices.

The importance of demand elasticity in determining how much of the tax a consumer will pay is illustrated in Panel III of Figure 6. With a sales tax, the supply curve shifts from S_1 to S_2. For relatively elastic demand, shown by D_1, equilibrium price increases from P_e to P_1. But with the more inelastic demand represented by D_2, the price increase or consumers' portion of the tax is much greater: The price increases from P_e to P_2. The consumer obviously

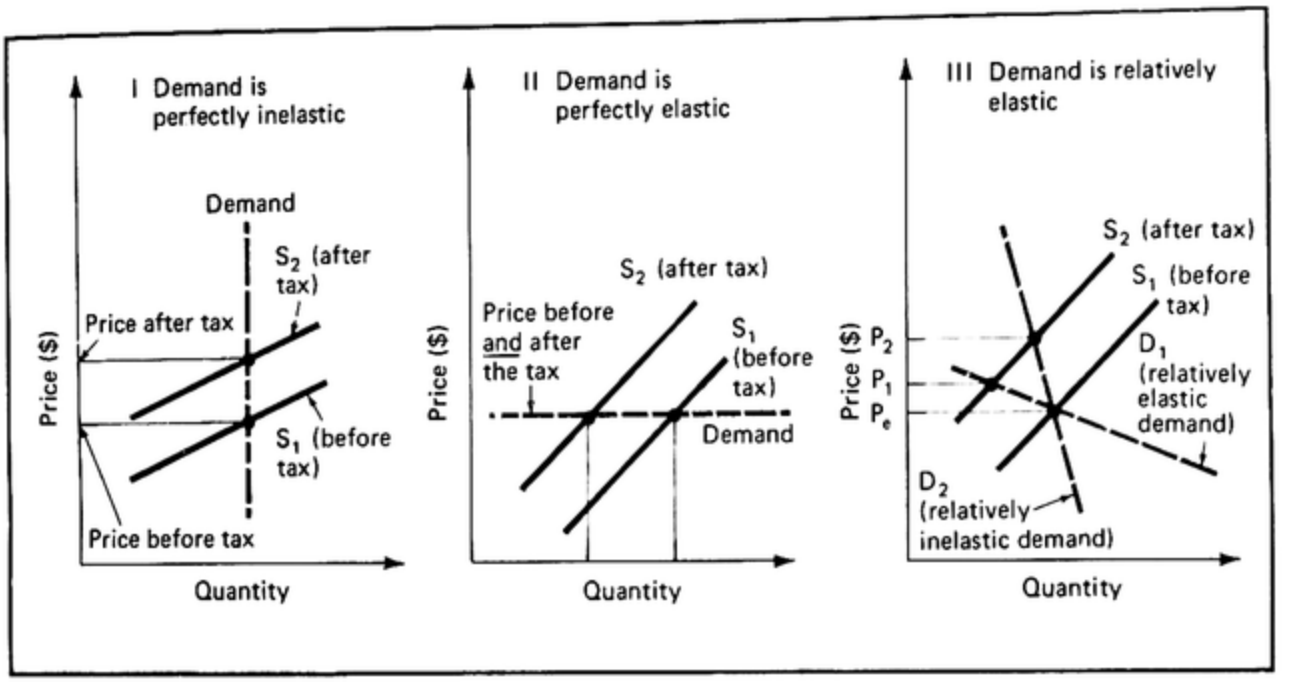


Figure 6 How demand elasticity affects the relative portions of a sales tax borne by the consumer and producer

If demand is perfectly inelastic, as it is in Panel I, a tax will raise the price of the good by the full amount of the tax and have no impact on quantity supplied. The consumer bears the full burden of the tax.

If demand is perfectly elastic, as it is in Panel II, the tax will have no impact on price. The full burden of the tax is borne by the producer.

If the demand is relatively elastic, as with D_1 in Panel III, the tax will cause a smaller increase in price to the buyer than it will if demand is relatively inelastic, as represented by D_2. In general, the more elastic the demand, the smaller the consumer's portion of the tax.

bears more of the tax the more inelastic consumer demand is.

However, the relative burdens of the sales tax are not determined solely by demand elasticity. Supply elasticity is also important. Figure 7 illustrates the various cases of supply elasticity. Panel I illustrates the case of perfectly inelastic supply. Here, the tax will not affect the supply curve, since the same quantity will be offered regardless of price. Since there is no change in price to the buyer, the producers bear the full burden of the tax. Panel II illustrates perfectly elastic supply. Here, the price increase equals the full amount of the tax. The producers receive the same revenue per unit after the imposition of the tax as they did before the tax, and thus bear none of its burden. Yet, the output falls to Q_1.

Most cases of supply elasticity lie between the extremes of perfectly elastic and inelastic supply. The more inelastic the supply is, other things being equal, the smaller the price increase will be. Therefore, the greater will be the reduction in the producers' revenue per unit, which represents the producers' share of the tax. The more elastic the supply, the greater will be the increase in price, and the smaller will be the producers' tax burden, represented by the decrease in producers' revenue per unit.

The effect of supply elasticity on the producers' share of the tax can be seen in Panel III of Figure 7. S_a is a relatively elastic supply curve, while S_b is a relatively inelastic supply curve. When the tax is imposed, both supply curves shift up by an amount equal to the tax. Look at the new equilibrium when supply is inelastic, with supply shifting from S_{b1} to S_{b2}. The producers' share of the tax is the difference between the before-tax revenue per unit of P_E and the after-tax revenue per unit of r_b. This reduction in producers' revenue per unit is shown by the segment $a-b$. When supply is more elastic, shifting from S_{a1} to

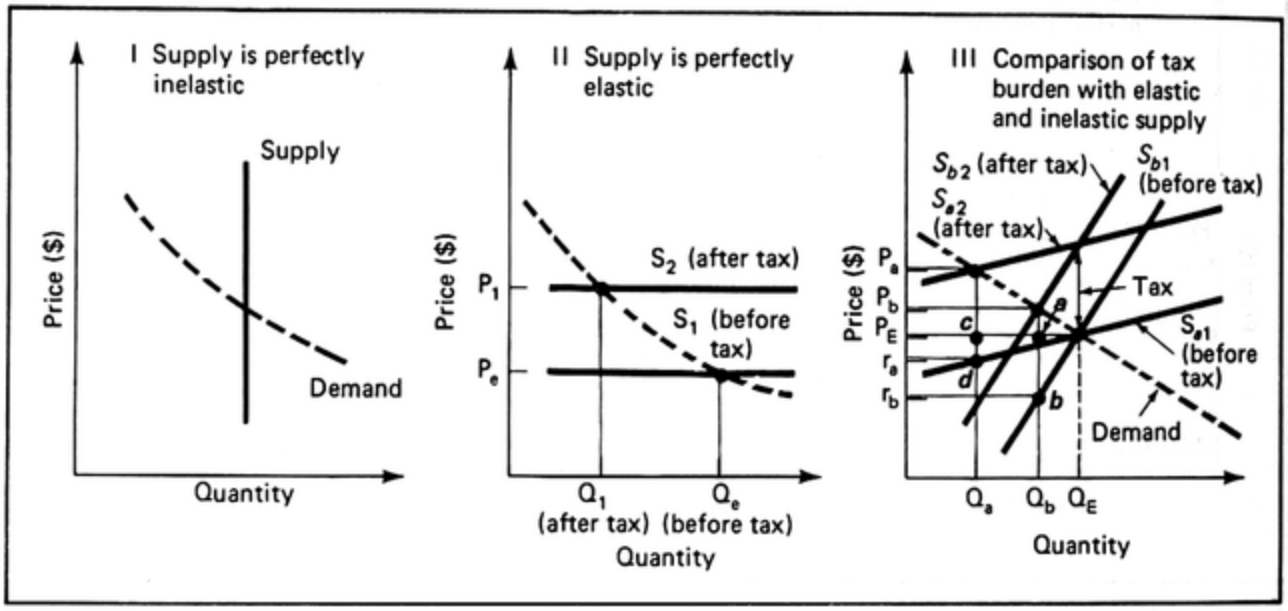


Figure 7 The influence of supply and demand elasticity on the effects of the sales tax

If supply is perfectly inelastic, as in Panel I, a tax will have no effect on supply since the same quantity of the good will be offered regardless of price. Here, because there is no change in price, the producers bear the full burden of the tax. If supply is perfectly elastic, as in Panel II, the price will increase by the full amount of the tax. The producers bear none of the tax burden. In Panel III, S_a represents a relatively elastic supply curve, while S_b represents relatively inelastic supply. With a tax, both supply curves shift up by the same amount, equal to the tax. For inelastic supply, (S_{b1} and S_{b2}), the equilibrium price rises from P_E to P_b. For elastic supply (S_{a1} and S_{a2}), equilibrium price rises by a larger amount, from P_E to P_a. The smaller increase in price when supply is relatively inelastic indicates that the producers are bearing a larger burden of the tax than with elastic demand, where the price rise and therefore the consumer share of the tax is larger.

S_{a2}, the producers' share of the tax—the difference between the before-tax revenue of P_E and the after-tax revenue per unit of r_a—is clearly smaller. It is shown by the segment c–d.

Combining both supply and demand elasticities, you should now be able to determine when the consumers would bear the biggest burden of the tax and, conversely, when the producers would bear the largest share. The consumers' portion of the tax will be larger, other things being equal, if supply is inelastic and demand is elastic. The producers' share of the tax will be larger when supply is inelastic and demand is elastic.

The general lesson is that taxes fall hardest on those who have the least flexibility (elasticity) in avoiding them. Governments have long followed this rule in deciding which goods to tax: ***Tax goods whose elasticities are low.*** If the demand for a good is inelastic, then even with the higher prices resulting from the tax, consumers will find it difficult to reduce their consumption of the good by very much. Since the government will only receive tax revenue from units of the good that are sold, it assures itself of the highest tax receipts if there is only a small drop in quantity sold as a result of the tax—that is, if demand is inelastic. Good examples of heavily taxed commodities with an inelastic demand are such habit-forming items as liquor and tobacco, and such necessities as gasoline today and salt in the Middle Ages.

Price controls

In all of the examples so far, a change in supply or demand resulted in market conditions automatically bringing price and quantity back to a new equilibrium or market-clearing level. Remember, though,

that there is nothing inherently fair or just, in a normative sense, about such equilibrium prices or quantities. They are simply the price-quantity combinations at which the market clears. Quantity supplied equals quantity demanded, so that there is neither an excess supply nor a shortage of the good.

There are cases in which the public believes that market forces will lead to a price that is unfairly high for consumers or unfairly low for producers. In these cases, the government may limit how high or low the market price may go. Now the market may be prevented from clearing because, by law, the price may be forced away from the equilibrium level. The following sections discuss the impact on a market of ***price controls.***

Price ceiling **A *price ceiling* is a legal limit on how much a supplier may charge for a particular good or service.** Price ceilings "protect" consumers by holding prices down. Often, price ceilings are applied to essential goods in short supply, such as food or housing. The rationale is that such goods should be available on an equal basis to all, not just to those who can afford to pay the high prices brought about by the short supply.

Such price ceilings have been attempted since time immemorial. There are many recent examples. General price controls were applied in the United States during World War II, the Korean War, and, to a smaller degree, in 1971–1974. Several cities, including New York, have long had rent controls, and many other cities have added them since 1965. Nearly every state has had a "usury" law limiting the rate of interest on loans. Natural gas prices have been controlled since about 1960, and oil prices were controlled during 1974–1981.

Panel I of Figure 8 shows what a price ceiling would do in a typical market. Without government interference, market forces would result in the equilibrium price P_E and the equilibrium quantity Q_E. What happens if the government imposes a price ceiling? If the price ceiling is above

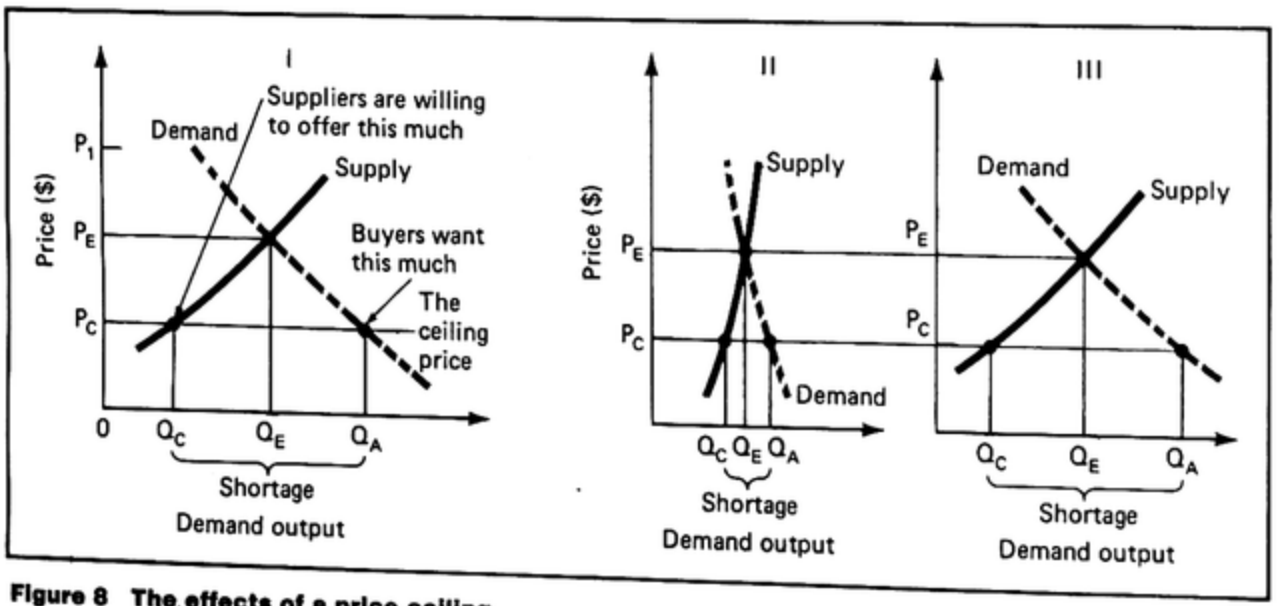


Figure 8 The effects of a price ceiling

In Panel I, representing a typical market, a price ceiling imposed *below* the equilibrium price prevents the market from clearing. At the ceiling price, consumers want more of the good than suppliers are willing to offer. A shortage of the good develops. Panel II shows the effects of a price ceiling when demand and supply are relatively inelastic. Panel III illustrates the effects of a price ceiling when demand and supply are relatively elastic.

the equilibrium price, such as at P_1, nothing happens. The price would not rise above P_1 anyway. The market can still clear at P_E and Q_E. To affect market, the price ceiling must be *below* the equilibrium price, such as at P_C. Now the price is prevented by law from rising to equilibrium, and the market cannot clear. At a price of P_C, consumers wish to purchase Q_A, while producers are willing to supply only Q_C. A physical shortage of the good develops, equal to the difference between Q_A and Q_C.

The effects of a price ceiling, then, are a lower price and a lower quantity supplied. Without the price ceiling, Q_E would have been supplied, rather than Q_C. Thus, while some consumers do benefit from the lower price, less of the good is available. The problem, then, is who is to get the good? Since price is no longer allowed to ration it, some other allocation system must be found.

One form of rationing is lining up or *queuing*. The people who can spend the most time waiting in line will then get the good. Or the government may impose its preferences by instituting *rationing*. Or, *sellers' preferences* may prevail. During gasoline shortages, for example, gas station operators often supply the scarce fuel only to their regular customers. Finally, bribes often guide the exchange of goods. Black markets—the illegal exchange of goods at prices above the price ceiling—often spring up.

How sharply a price ceiling cuts back on the quantity supplied will vary with the elasticities of supply and demand. Try drawing a price ceiling for a good with a relatively high inelastic supply and demand, and for a good with a relatively high elastic supply and demand. This is done in Panels II and III of Figure 8. You can see that if both supply and demand are highly inelastic, as in Panel II, then the lower-than-equilibrium price will not cause much of a reduction in quantity supplied, or much of an increase in quantity demanded. In that case, the shortage resulting from the price ceiling will be small and, perhaps, easily managed. But if both supply and demand are highly elastic, as in Panel III, a price ceiling will cause a severe shortage, and allocating the quantity will be a serious problem.

Once again, you are in trade-off territory. The benefits of the lower price, especially to the most needy groups of people, must be weighed against the costs from having a lower quantity supplied.

Price floor **A *price floor* is a legal limit on the minimum price that a supplier may charge for a particular good or service. Price floors benefit the suppliers of a good or service.** Agricultural price supports are one form of price floor that has been common since the 1930s. You saw earlier in the chapter that inelasticity of demand for agricultural products, coupled with the variability of the size of harvests, causes farm incomes to fluctuate. Agricultural price supports have sought to stabilize farm prices, given the unpredictable nature of harvests.

Figure 9 shows what a price floor would do in a typical market. Without government interference, market forces would result in an equilibrium price of P_E and an equilibrium quantity of Q_E. What happens if the government imposes a price floor of P_1 *below* the equilibrium price of P_E? Nothing. The price would not fall below P_1 anyway. To affect the market, a price floor must be *above* the equilibrium price, such as P_F in Figure 9. Now the price cannot fall to the equilibrium level, and the market cannot clear. At a price of P_F, the buyers wish to purchase Q_A, and the suppliers wish to offer Q_F. Thus, a *surplus* or excess supply of the good occurs, equal to the difference between Q_A and Q_F.

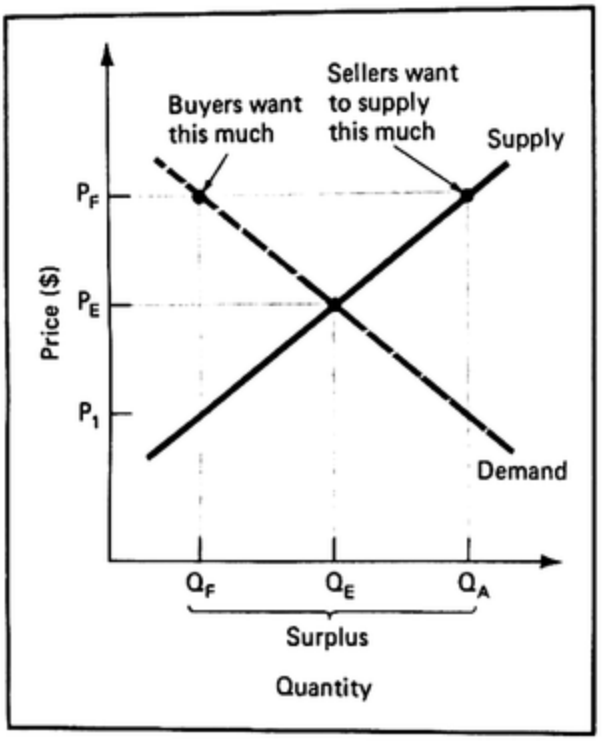


Figure 9 The effects of a price floor

For agricultural price supports, the surplus is crops that must be stored. From the 1930s to the present, farm price supports have caused large surpluses to build up in hundreds of storage sites around the country, at a cost of many billions of dollars.

Controls on quantity

So far, you have seen that price controls always affect the quantities supplied and demanded. Sometimes, however, a government tries to control quantity directly. As you might expect, this affects prices.

The most obvious and common form of ***quantity control*** is a flat prohibition of a good, such as marijuana. Suppose that if there were no restrictions on growing, selling, and using marijuana, the supply and demand curves would be represented by S_1 and D_1 in Figure 10. The market equilib-

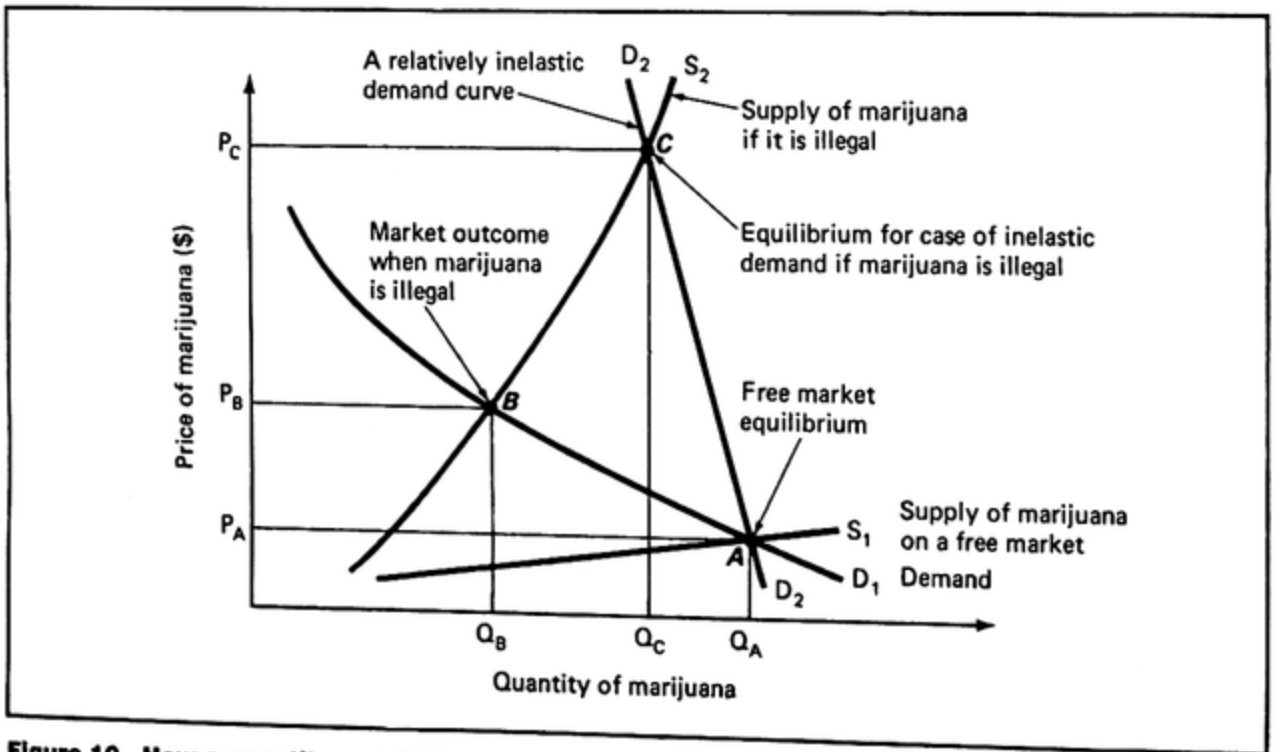


Figure 10 How a quantity prohibition may affect the price and quantity of marijuana

The prohibition is formally meant to yield a supply of zero. But S_2 results, causing both a significant reduction in quantity and a significant increase in price. The more inelastic the demand, the smaller the decrease in quantity and the larger the increase in price. You can see this by comparing the changes resulting in the case of relatively elastic demand D_1 (Q_A to Q_B and P_A to P_B) with the changes resulting in the case of relatively inelastic demand D_2 (Q_A to Q_C and P_A to P_C).

rium for this good would result in P_A and Q_A. A totally effective prohibition on the use of marijuana would mean that supply would be zero at every point—simply a vertical line coinciding with the Y axis.

In practice, even though it is illegal, quantities of marijuana are bought and sold covertly. But the supply situation is completely different from that shown by Curve S_1, since supply is now much more costly. Efficient local farming of marijuana is largely prevented, so that supplies are imported, which greatly adds to the expense. Suppliers must also expect an occasional confiscation, fines, and jail terms as part of the cost of doing business. Supply Curve S_2 might now apply, illustrating that quantities will be available only at higher prices.

If demand remains at D_1, the new equilibrium of P_B and Q_B shows just what you would expect—a higher price and a lower quantity. If the demand for marijuana is highly inelastic, as shown by D_2, the prohibition on marijuana results in a staggering increase in price. At the equilibrium represented by C, much more money is being spent on marijuana than at the original equilibrium of Point A, and most of this revenue is going to criminals.

Yet, because the purchase of marijuana is now a criminal offense, many purchasers drop out of the market. They are not willing to risk a fine or imprisonment to obtain the drug. The ultimate impact on equilibrium price and quantity of these duals shifts—the decrease in both demand and supply—depends upon the relative size of the supply and demand changes. The quantity exchanged will clearly decrease, as an effect of both the decrease in supply and the decrease in demand. But the impact on price is not so easy to determine. The decrease in supply would cause the price to rise, while the decrease in demand would cause the price to fall. The actual impact on price cannot be determined unless the relative sizes of both the supply and demand shifts and the elasticities are known. The experience in the United States seems to indicate that the price for marijuana is higher than it would be if it could be openly exchanged under competitive conditions.

Measuring supply and demand

You have now seen how useful supply and demand can be in explaining the market outcomes in a wide variety of situations. Even without knowing much actual data, you can use supply and demand analysis to show which way price and quantity will move.

But to make precise estimates of the new equilibrium prices and quantities is much more difficult. That is because demand and supply are not easy to measure. Consider why.

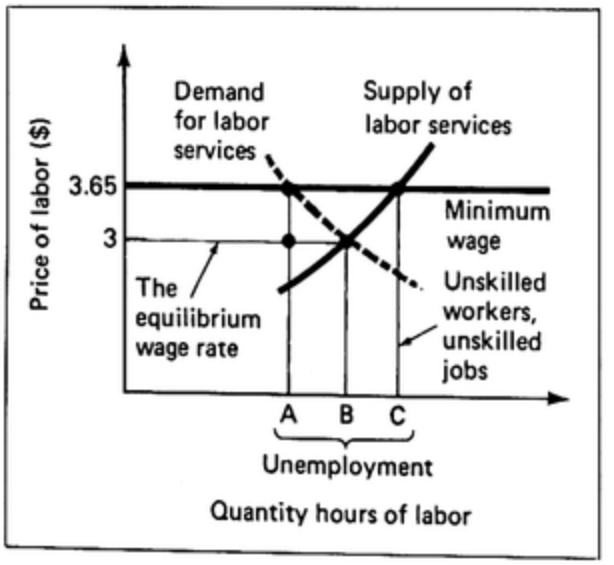
In any given period, one price-quantity combination prevails in a market. That actual price and quantity shows one common point on both the demand and supply curves for that market. That is all that can be measured directly. The price-quantity combinations that are known from past periods may have come from supply and demand curves that had different positions. In fact, demand and supply schedules often do shift from one period to the next.

If the demand curve remained stationary while only the supply curve shifted, a demand curve would be traced out. If the supply curve stayed stationary and only the demand curve moved, then a supply curve would be traced out. These two cases are shown in Panels I and II of Figure 11.

Yet, both curves are often shifting at the same time. The series of points showing price-quantity combinations in successive periods is then simply a scattering of

Example of a Price Floor: How the Minimum Wage Creates Unemployment

Suppose the equilibrium wage is $3 an hour, but the government sets the minimum wage at $3.65 an hour. What happens? As the figure shows, the quantity of labor demanded decreases (from B to A), while the quantity supplied increases (from B to C). In other words, more people want to work even though there are fewer jobs for them. The resulting unemployment and the size of workers' incomes depend on the elasticity of the demand for labor. When it is inelastic, unemployment will be less and income will rise. When it is elastic, there will be more unemployment, and income will fall.



dots that proves nothing, as is illustrated in Panel III of Figure 11. The time series of price-quantity values will never allow us to trace out both the supply and demand curves, and only rarely will it allow us to trace out one curve with great confidence.

Nevertheless, economists have developed econometric methods to identify the true supply and demand curves from real-world data. Agricultural products have been most intensively analyzed, but other goods have also been studied. These measured demand and supply elasticities are estimates, not definitive values. Yet, they often show the elasticities with some accuracy.

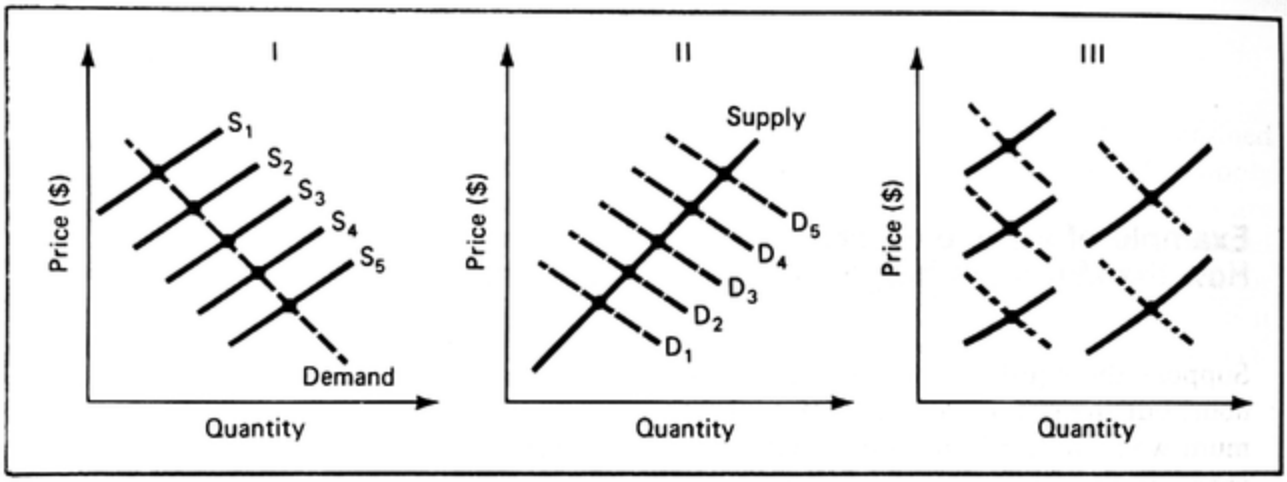


Figure 11 Shifts of supply and demand curves

In Panel I, the demand curve remains stationary while the supply curve shifts, tracing out the demand curve. In Panel II, the supply curve remains stationary while the demand curve shifts, tracing out the supply curve. In Panel III, both the supply and the demand curves are shifting. The resulting scatter of points tells nothing about either supply or demand.

Table 1 Measures of price elasticities of demand for selected goods and services in recent years

Good or Service	Short-Run Price Elasticity	Long-Run Price Elasticity
Gasoline (transportation only)*	−.20	−1.50
Automobiles and parts	−.72	−1.10
Furniture and household equipment	−.27	−.90
Food and beverages	−.23	−.58
Clothing and shoes	−.20	−.33
Gasoline and oil	−.07	−1.03
Housing	−.006	−.37
Transportation	−.09	−.55
Tobacco products	−.46	−1.89
Shoes	−.73	−1.21
Jewelry and watches	−.41	−.67
Toilet articles and preparations	−.20	−3.04
Automobile repair	−.40	−.38
Radio and TV repair	−.47	−3.84
Movies	−.87	−3.67
Theater and opera	−.18	−.31

	Price Elasticity
Bread†	−.15
Beef†	−.64
Lamb and mutton†	−2.65
Eggs†	−.32
Hospital and physician service‡	−.10
Electricity (long-run commercial and residential)*	−.88

Sources:
*J. M. Griffin, *Energy Conservation in the OECD, 1980–2000* (Cambridge, Mass.: Ballinger, 1979).
†P. S. George and G. A. King, *Demand for Food Commodities in the United States with Projections for 1980* (Berkeley: University of California, 1971).
‡Joseph P. Newhouse and Charles E. Phelps, "New Estimates of Price and Income Elasticities of Medical Care Services," in Richard N. Rosett, ed., *The Role of Health Insurance in the Health Services Sector* (New York: National Bureau of Economic Research, 1976).
All others: H. S. Houthakker and Lester D. Taylor, *Consumer Demand in the United States: Analyses and Projections* 2nd ed. (Cambridge, Mass.: Harvard University Press, 1970).

Table 2 Estimated elasticities of supply for selected agricultural products

Commodity	Short-Run Price Elasticity	Long-Run Price Elasticity
Green lima beans	0.10	1.70
Cabbage	0.36	1.20
Carrots	0.14	1.00
Cucumbers	0.29	2.20
Lettuce	0.03	0.16
Onions	0.34	1.00
Green peas	0.31	4.40
Green peppers	0.07	0.26
Tomatoes	0.16	0.90
Watermelons	0.23	0.48
Beets	0.13	1.00
Cantaloupes	0.02	0.04
Cauliflower	0.14	1.10
Celery	0.14	0.95
Eggplant	0.16	0.34
Kale (Va. only)	0.20	0.23
Spinach	0.20	4.70
Shallots (La. only)	0.12	0.31

Source: M. Nerlove and W. Addison, "Statistical Estimation of Long-Run Elasticities of Supply and Demand," *Journal of Farm Economics*, November 1958.

A selection of measured elasticities is shown in Tables 1 and 2, for both demand and supply. Note the wide range of the values, from low to high elasticities. Compare these estimates with the values that you would expect for each type of good. Note that *long-run* demand elasticities are all higher than the corresponding short-run elasticities, for reasons explained in Chapter 4. Note, too, that such basic items as food and clothing have, as expected, lower elasticities than luxuries such as toilet articles or movies.

In sum, the logic of supply and demand analysis can show the direction in which the economic outcome will go. But it often cannot predict exactly *how far* the changes will go or *where* the new equilibrium will be. Economists can usually make valid (though imprecise) predictions, even if they can rarely settle an issue completely.

Summary

This chapter shows how supply and demand analysis can be used to examine various market situations. The major points of the chapter are:

1. All markets are influenced by changes in supply and demand conditions. When supply and demand shift, the resulting changes in market prices and quantities will depend crucially on the elasticities of supply and demand.
2. In some cases, market changes are triggered by events within the market itself.
3. In some cases, market changes are triggered by government intervention in the market process.
4. The burden of a *sales tax* is usually borne by both the consumer and the producer. Their relative shares are determined by the supply and demand elasticities.
5. A *price ceiling* is a legal limit on the maximum price that a supplier may charge for a good or service. Such ceilings are intended to protect consumers by holding prices down. If the market price is held below the equilibrium level, a shortage of the good will occur, since quantity supplied will be less than quantity demanded.
6. A *price floor* is a legal limit on the minimum price that a supplier may charge for a good or service. Such price floors are meant to benefit the supplier. If the market price is held above the equilibrium price, an excess supply of the good will result, since the quantity demanded will be less than the quantity supplied.
7. Controls on quantity are often a flat prohibition, such as on marijuana. The effect of such a quantity control will usually be higher prices and lower quantities exchanged. If demand for

the good is highly inelastic, the increase in price will be enormous, and the decrease in quantity will be small.

8. If the demand curve is stationary while the supply curve shifts, a demand curve will be traced out. If the supply curve is stationary while the demand curve shifts, a supply curve will be traced out. If both curves are shifting at the same time, the series of points representing price and quantity combinations from one period to the next is simply a scattering that shows nothing.

Key concepts

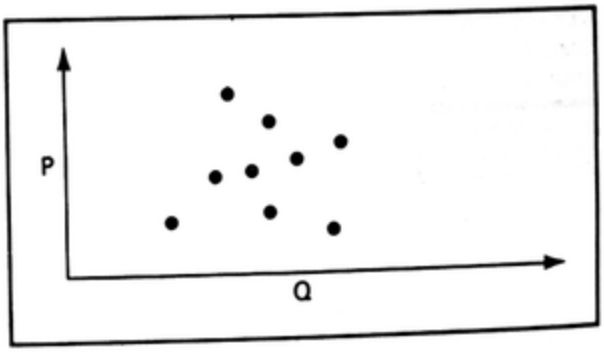
Sales taxes

Price controls
- price ceiling
- price floor

Quantity control

Questions for review

1. a. Start with a basic supply and demand graph. On your diagram, show the effects of the enactment of a per unit sales tax. Indicate both the consumers' and producers' share of the tax.
   b. List some goods for which you would expect either the consumers' or producers' share of the tax to be relatively large. Explain.
2. Using supply and demand analysis, explain why agricultural price supports are often accompanied by quantity restrictions, such as a limit on the number of acres planted.
3. Suppose that observations of actual price-quantity combinations in a particular market show the following scatter of points:



Would it be reasonable to conclude that this is one good for which quantity demanded is not influenced by price?

6

Individual Demand

As you read and study this chapter, you will learn:

- the basis of rational consumer choices
- the law of diminishing marginal utility
- the consumer's equilibrium conditions
- the determinants of demand
- how to derive market demand from individual demand
- the underlying meaning and limits of demand analysis

The Scene: Supper time in the Broomfield family dining room.

Richard: Dad, can we get a better car? Our car is too old.

Mr. Broomfield: Maybe we could, but it would be expensive. Then we'd have to eat out less, buy fewer new clothes and records, and take shorter vacation trips.

Mrs. Broomfield: Maybe we could sell Dad's motorboat—

Mr. Broomfield: Wait a minute, wait a *minute.*

Karen: Why, Mom? We need all of these things.

Mrs. Broomfield: We can't afford them all on our income. Having more of some means that we'll have to get by with less of others.

Karen: Oh, you mean that we must adjust the quantities as we optimize our spending to reach equal marginal conditions among all goods?

Mr. Broomfield: That's economic jargon. We're talking about real life.

Karen: But the two are the same!

Indeed, like millions of other families, the Broomfields are applying their preferences, considering prices, and continually adjusting toward their best combination of purchases. Will they sell Dad's motorboat to buy the new car Richard wants? Read on to learn the economic basis for their—and everyone's—decisions.

To show you that, we will take you behind the simple line of the market demand curve. You will learn how demand arises both from people's individual preferences and from the money they have to spend. You have already seen how consumer preferences are summed up in the market demand curve. Now we narrow the focus of our economic microscope to the individual consumer or family as it decides how to divide its income among all the goods that are available in the economy. The demand curves presented in this chapter do not show the total amount of the good purchased in the whole market at various prices. Rather, each one shows the quantity of a good that an *individual* will buy at various prices over a certain time.

Since the market demand curve is derived by adding up individual demand curves, the two types of demand schedules have much in common. Above all, the market demand curve slopes downward, as you know from Chapter 4. That occurs because ***individual people's demand curves*** **—from which market demand is derived—** ***slope down.*** You will soon see why that is true, as you learn how consumers endlessly adjust their choices as prices vary.

The starting point is a pair of basic facts about people's preferences. First, people are *alike.* They all need food, shelter, clothing, and a few other items. But people are also *different.* They insist on liking different things. Jones loves grand opera; Thomas hates it. Cruz loves pickles and drag racing; Newman hates them. Steingut loves her 20th-floor high-rise apartment in Chicago; Wyatt yearns for a ranch under Wyoming's big sky. You yourself are a bundle of specific likes and dislikes: Hawaiian shirts? Elvis Presley records? Pinstripe suits? Classical ballet? Diet foods? Cigarettes? Many of these differences are felt intensely.

Somehow an efficient economy needs to accommodate these differences, offering Jones his opera, Cruz her pickles and drag racing, and Steingut her apartment. For that to happen, these preferences must be expressed. In a market economy, the preferences are embodied in *individual demand.* This chapter shows the causes and nature of that demand.

First, we present the basic concepts. Declining marginal utility is the key concept in rational consumer choices (as we previewed in Chapter 2). Next, we show how the rational consumer reaches equilibrium in choosing among many goods. Then we explain several technical features of demand and derive the market demand curve. Demand analysis is not perfect or all-powerful, but the second section discusses its great underlying strengths.

The analysis of utility and demand

It is best to begin by learning the consumption patterns in the economy as a whole, as summarized in Table 1. There are about 75 million U.S. households (families, singles, and others), with a yearly average of $15,000 in consumption purchases. Altogether, the largest share of personal spending (21.3 percent in 1979) goes for food, beverages, and tobacco. Clothing, transportation, and housing-related expenditures are also important.

These averages mask sharp variations among individual consumers. Some families have modest housing but eat expensive

Table 1 ***Where the money goes: A typical U.S. family***

	The Share of Personal Spending			Amount for a Typical Family
	1950 (%)	1965 (%)	1979 (%)	1979
Food, beverages, and tobacco	30.3	24.9	21.3	4,159
Clothing, accessories, and jewelry	12.3	9.4	7.8	1,523
Personal care	1.3	1.8	1.3	254
Housing	11.3	15.2	16.0	3,124
Household operations	15.2	14.2	14.5	2,831
Medical care	4.7	7.0	9.7	1,894
Personal business	3.4	4.6	5.4	1,054
Transportation	13.2	13.5	14.1	2,753
Recreation	5.8	6.0	6.7	1,308
Other	2.4	3.4	3.2	625

Source: *U.S. Statistical Abstracts*, 1978, p. 431; 1980, p. 442

food. Some families travel a lot, others not at all. And within each category (such as food), many families vary even more sharply in their specific choices (e.g., among turnips, cheeseburgers, lobster, and ribs). The averages only hint at the full variety of consumer spending.

You can also see from Table 1 that spending patterns shift over time. Between 1950 and 1979, for example, the share spent on both food and clothing dropped by nearly one-third. Meanwhile, the shares spent for housing rose by nearly half, and medical care's share more than doubled. These large changes reflected many causes, including changes in consumers' tastes and relative prices. Income elasticities also were important as average incomes rose. Families apparently moved from relatively inferior goods (food, clothing) to such high-income-elasticity goods as housing and medical care.

The table's neat columns of percentages and numbers are a shorthand way of summarizing the choices of the U.S. population. Those choices reflect people's real preferences and budgets as they take hold in real markets. The crucial task is to analyze how these consumer choices are made.

Rational choices by consumers

Some consumers are single people who live alone and make their choices strictly by themselves. But most live in families or other groups, sharing in the household choosing and spending. The same logic, however, applies to all of them. By "the consumer," economists mean any decision unit, with one or several participants. (The Broomfield family at the start of this chapter is one such unit, hammering out its preferences and choices.)

Rational choices by a consumer will lead to the highest possible level of satisfaction, given the amount of income that the consumer has to spend. The economist's technical term for consumer satisfaction is ***utility.***

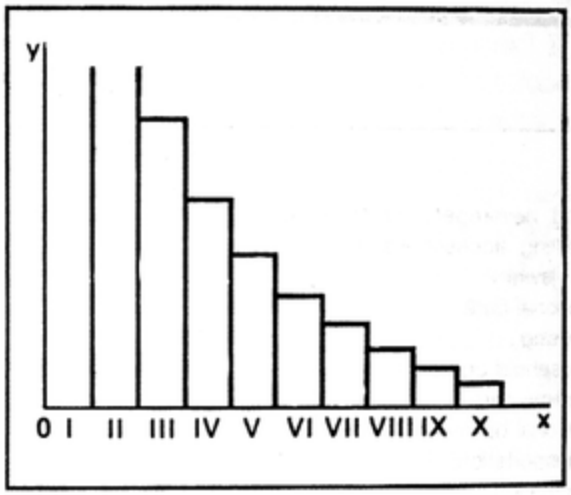
Economists assume that more utility is better than less. Each good can give some utility, but there are many goods to choose among. The consumer must select

Utility and Marginal Analysis

The concept of utility originated with an eccentric English philosopher, Jeremy Bentham (1748–1832), and his group of fellow "utilitarians" in the 1780s. Their goal was to improve the welfare of individual people, defined as their utility or satisfaction.

Utility acquired precise meaning only in the 1870s, with the neoclassical marginal utility analysis of William Stanley Jevons (1835–1882) and others. By stressing utility, Jevons put demand on par with supply as the two determinants of value. In his *Theory of Political Economy*, his first diagram of declining marginal utility—reproduced here—is much like those that economists still often use.

Jevons' pioneering effort, together with work by Carl Menger, Leon Walras and Alfred Marshall formed much of the neoclassical revolution that established marginal analysis as the core of economics. Marshall's main contributions included not only the concept of elasticity, but also the advanced analysis of consumer surplus and economic rent. His massive *Principles of Political Economy* both assembled and refined the body of new neoclassical thought for many decades of later economists.



Jevons' marginal utility diagram

The x on the horizontal axis is food. The y axis shows the degree of utility. The first two units of food give utility that is not specifically measured because "those portions of food would be indispensable to life, and their utility, therefore, is infinitely great." Diminishing marginal utility is readily apparent.



WILLIAM STANLEY JEVONS

the array of goods that will result in the greatest amount of utility possible, within the limits of his or her budget. *The goal of the consumer, in short, is to maximize utility subject to a budget constraint.*

Because utility is only a state of mind, each person has to make his or her choices independently. The utility-maximizing choices can't be arranged from outside by someone else. Indeed, your preferences are not just an economic datum. They express much of your whole personality. Such intensely personal inner conditions are ultimately the wants that the economy ought

to service. Therefore, when economists focus their analysis of demand on *utility*, they are trying to expose how some of life's most important decisions are made.

Note also that utility cannot be measured accurately, as by some electronic meter. Pleasure is measured in feelings and attitudes, not according to rigid formulas. No consumer will actually use this chapter's precise concepts and diagrams to make consumption choices. For the same reason, no bluejay needs to study aeronautical engineering to fly. Bluejays can fly without advanced training, and consumers can make reasonably wise decisions without reading this chapter.

Yet the analysis of demand is nonetheless valid. ***The mass of consumers usually reach choices that are much the same*** as if ***they were applying the precise analysis of utility that follows.*** This *as if* hypothesis is crucial and powerful. It enables us to see order in the vast flow of human decisions, many of them made seemingly in haste and without planning. Even the person deciding impulsively to buy a shirt, a book, or a used car is probably behaving *as if* the choices were rational. In this light, the economist's task is to derive a clear logical analysis of the spending decisions that a consumer makes intuitively.

Diminishing marginal utility

In Chapters 4 and 5 we showed that market demand curves slope down: Price and quantity are inversely related. The explanation for that fundamental condition centered on income and substitution effects. Now our focus shifts from markets to the individual consumer. **Individual demand curves also slope down because of the *law of diminishing marginal utility.***

Total utility Consider any economic good —hamburgers, for example. A standard hamburger can provide utility. Economists speak of utility as a degree of pleasure that arises from consuming specific goods. They do not expect to measure it directly, in units of satisfaction, or "utils." But economists suppose that utility is experienced as a magnitude that can be illustrated, as in Figure 1. Accept this convention for now; the meaning of utility will be discussed again later in the chapter. Returning to the hamburgers: As you eat one and then another, the ***total utility*** you derive from hamburgers may rise, as is illustrated by the curve at the top of Figure 1 and the

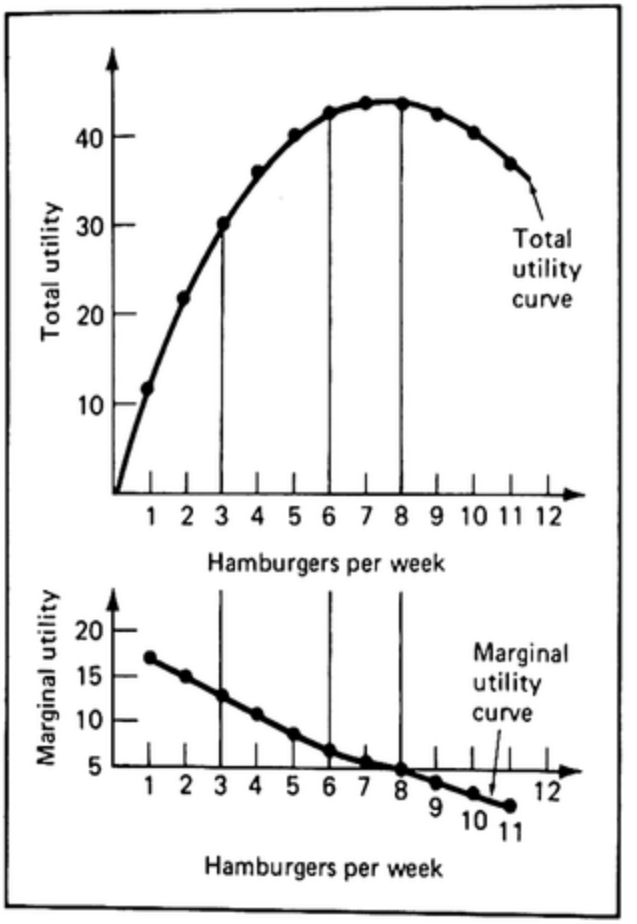


Figure 1 Marginal utility is related to total utility

Total utility rises as consumption of a good increases, but the rise tapers off. The rate of rise is shown by the slope of the total utility curve, which is the change in total utility for each added unit of the good consumed. Note that this is precisely the definition of *marginal utility*. Thus, while total utility rises, marginal utility—the slope of the total utility curve—declines. At eight hamburgers per week, marginal utility has sunk to zero: No more hamburgers are wanted at all. Beyond that level, each hamburger reduces total utility. All of this illustrates the law of diminishing marginal utility.

Table 2 Total and marginal utility: An illustration

Number of Hamburgers per Week	Total Utility	Marginal Utility
0	0	—
1	12	12
2	22	10
3	30	8
4	36	6
5	40	4
6	42	2
7	43	1
8	43	0
9	42	−1
10	40	−2
11	37	−3

numbers in Table 2. There we measure the total utility from hamburgers eaten per week.

Total satisfaction rises as you go through the first six hamburgers per week, but the rise tapers off. The first hamburger provides 12 "utils," while the second contributes only 10 units to total utility. Since both amounts are positive, total utility still rises. But by the sixth hamburger, the marginal utility is only 2, and when you reach the seventh hamburger of the week, you feel that the next one won't make you happier at all. And from the eighth hamburger on, you endure decreasing total utility. Each added hamburger makes you feel worse than before and more worried about gaining weight.

This rise and then decline of total utility is virtually a universal pattern. It holds for every ordinary economic good and for every ordinary consumer. It is true even though each person's unique preferences give a unique, specific total utility curve. It also remains true even though utility cannot be precisely measured. The total utility curve merely shows formally how a person feels about the various amounts of the good.

Marginal utility But it is on ***marginal utility*** that economists focus to show the pattern even more clearly. ***Marginal utility is the change in total utility from adding one more unit of the good.*** It is shown in the bottom part of Figure 1, lined up precisely below the total curve. The horizontal axis—the amount of the good—is the same for both diagrams. The marginal utility of that third hamburger is therefore 8, and that value is shown in the marginal utility curve straight below.

The marginal utility curve shows with special clarity how consumption levels affect satisfaction. Notice that marginal utility declines throughout. The decline reflects the basic fact that the first unit is the best; marginal utility declines as consumption increases. By definition, marginal utility is still positive as long as total utility rises. That occurs in Figure 1 up to the seventh hamburger. But when total utility peaks, marginal utility is zero and heading into the negative range below the horizontal axis. Marginal utility is declining throughout the diagram, but it actually becomes negative as the quantity increases from eight to nine per week.

Negative marginal utility means *displeasure; it is called disutility.* Therefore, you find the ninth hamburger (and all additional ones) *bad,* not good. The marginal utility curve shows this crossover point—between added pleasure and added displeasure—even more clearly than the total utility curve above it.

The logic of the analysis can be confirmed by comparing this diagram with your own feelings about a number of goods that you regularly consume—eggs, blankets, and shampoo, for example. You find them useful goods, but their marginal utility declines as you use more of them. And at some level, such as 8 eggs per day, 5 blankets on your bed, and 12 bottles of shampoo per month, you wouldn't really want any more of them, even if they were

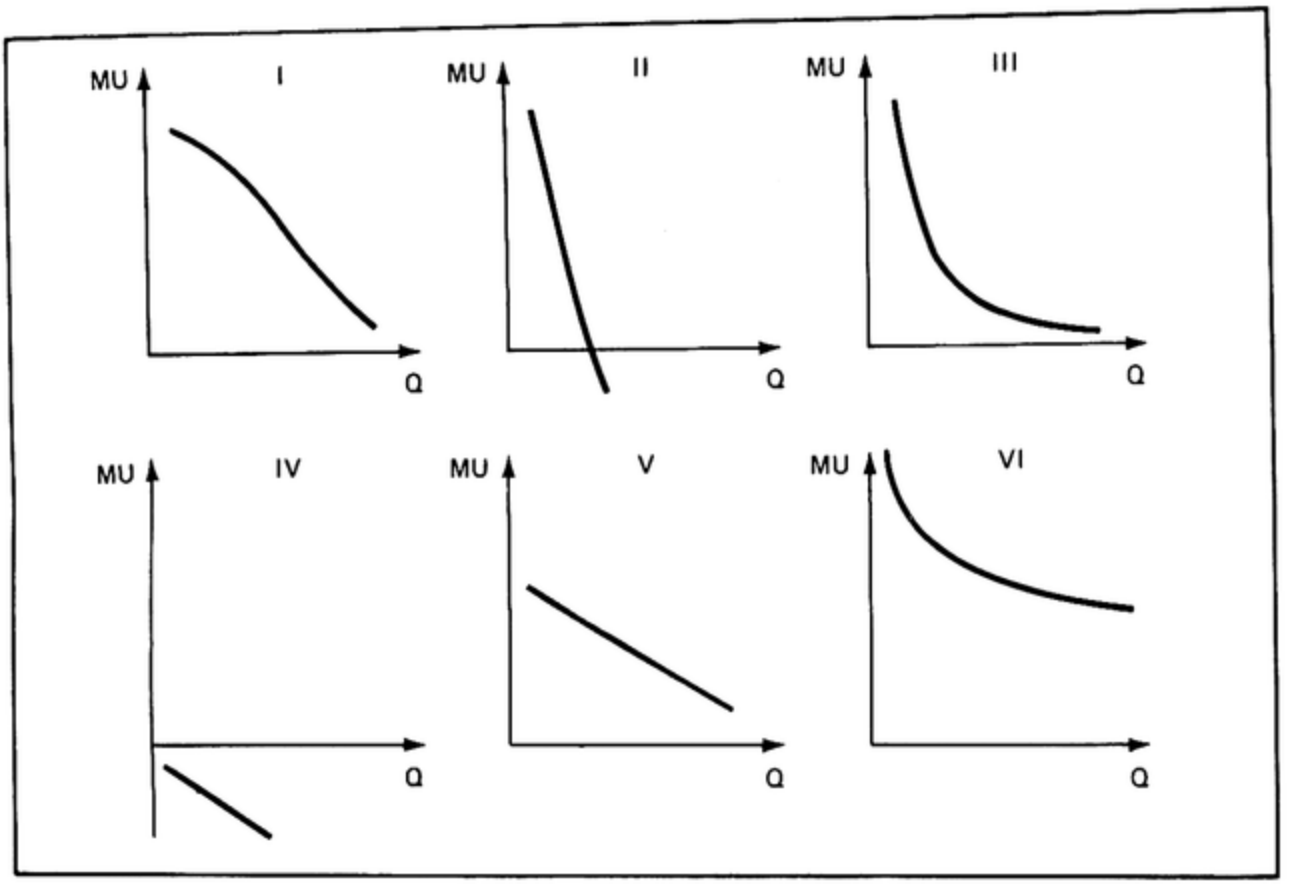


Figure 2 A few of the many forms that marginal utility curves can take

given to you free. The lesson: The marginal utility of each good does decline, and you judge rather clearly where the marginal utility curve cuts zero. With careful thought, you can locate the crossover point for every good you use.

You can also make these rough judgments from what you and your friends *do*. If you or a friend drank a third glass of orange juice with breakfast, it probably had positive marginal utility. If you see people taking seconds of salad or apple pie, their marginal utility was positive. But when you decline a fourth hamburger—or any other offer—your marginal utility is zero or negative.

Variety Because people's preferences vary, their marginal utility curves differ. For example, some people love liver; others won't touch it. Having a .45-caliber pistol at hand makes some people happy; others would hate it. No two people will emerge from a supermarket with exactly identical carts full of groceries, and often their choices are radically different. The variety of human preferences is an important fact of economic life.

Utility analysis makes this variety clear. As Figure 2 illustrates, marginal utility curves may be high or low. They may slope down sharply or be nearly flat. They may have all kinds of bends and twists, *as long as they do slope down*. Even for the greatest delights, repetition dulls the pleasure.

The curves may be entirely in the negative range, as is Curve IV in Figure 2. Each person has many dislikes, and such "bads" are shown by negative marginal utility. The person considers them a form of garbage, even though other people may love them.

It is helpful to practice drawing marginal utility curves for 10 or 20 different goods, getting a feel for different preferences and cases. Include some goods that

you really don't like at all, as well as some of your favorites.

At any rate, marginal utility is the bedrock of consumer choice. Economists show marginal utility by curves that (1) slope down, (2) can take many shapes, and (3) can lie entirely in the negative range.

Individual demand curves: Preferences and income

The marginal utility curves illustrate states of mind, which are based on inner preferences. They are independent of income and the prices of the goods. They are also private and hard to express precisely. But they govern what people actually do in the marketplace, and that is what matters for the economic process. *These consumer actions are embodied in individual demand curves.* Each curve relates price and quantity for one good and one person. It shows how much a person will buy at each alternative price.

Since the individual demand curves are what make up market demand curves, many of the influences on market demand that we presented in Chapter 4 will also influence individual demand: the price of the good, the price of other goods, expectations about future prices, income, and preferences. (Income distribution and total population are the two influences on market demand that are not relevant for an individual demand curve. After all, the amount of a good you choose to consume does not depend on how many other consumers there are or what their incomes are.)

A person's demand curve is derived by allowing only price to vary while keeping the other influences fixed. **At each price the consumer will choose to buy a specific amount of the good. Varying the price generates a series of such points. Together, those points make up the demand curve.**

The demand curve will slope downward. This reflects the law of diminishing marginal utility: Marginal utility declines as additional units of a good are consumed. If additional units of a good add less and less to your satisfaction, you will only purchase these units if the smaller increases in satisfaction are matched by lower prices. The price reflects what you sacrifice of other goods by not using the money to buy them. If the tenth unit gives you less satisfaction than the first unit of a good, you will not be willing to pay as much for the tenth unit as you were for the first.

Those choices rest closely upon personal preferences, for it is preferences that move people to act. Therefore, the demand curves have the same general shapes and slopes as the marginal utility curves. If marginal utility is high at first but then slopes sharply down, the demand curve will also. If your marginal utility for a good drops off rapidly, it will take fairly large price decreases to persuade you to consume more. If your marginal utility for a good drops off slowly, then it will take smaller decreases in price to persuade you to purchase more of the good. Both the demand curve and the marginal utility curve will cut the horizontal axis at the same quantity. And for a "bad," with a marginal utility curve that is negative throughout, a person's demand curve will also lie entirely below the horizontal axis.

Preferences are crucial. *But equally important is the consumer's income,* which controls how much he or she can afford to spend. The consumer's income affects the vertical height of the demand curve. Wants must be backed up by money to be expressed in market choices. Suppose that both Smith and Jones have similar conditions of utility as shown in marginal utility curves, but Smith is very rich while Jones has only $5,000 income per year. Smith's

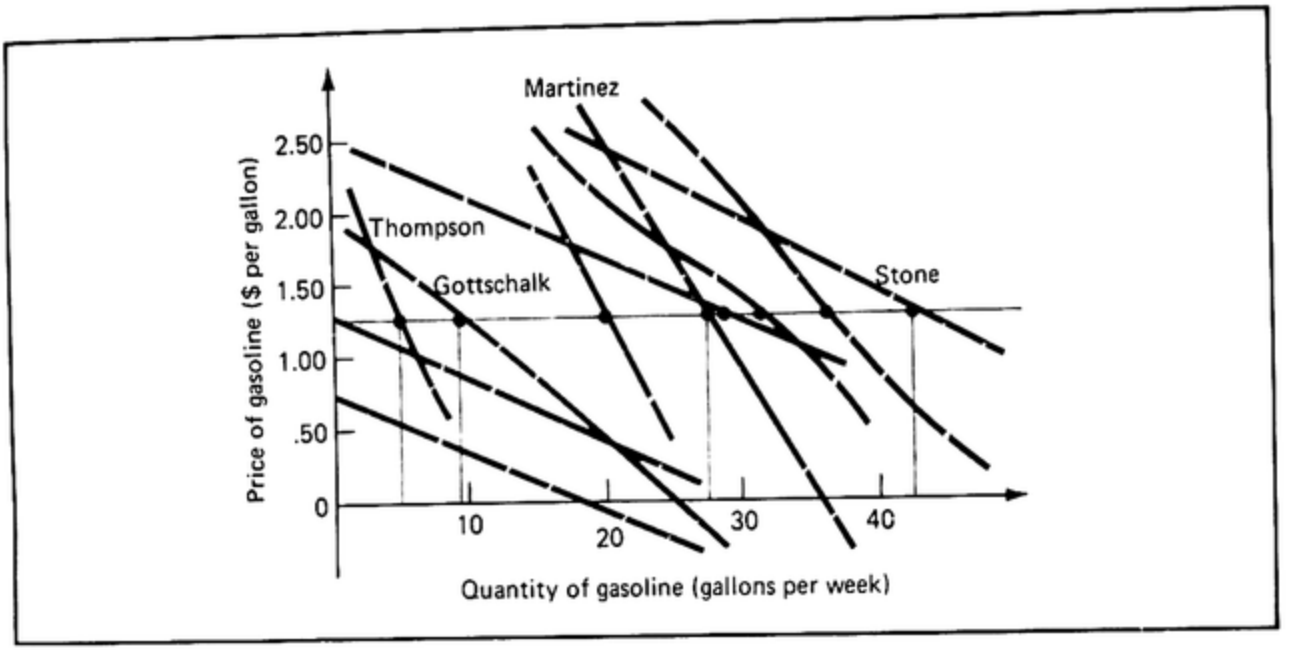


Figure 3 At any given price, individual demand curves show what quantities will be taken
At $1.25 per gallon, the gallons consumed by these ten people are as shown. Some people consume none at all; others take as much as 42 gallons per week.

demand curves for most goods will lie far above those of Jones. Indeed, even if Jones's longing for the goods were far more intense than Smith's, Jones simply could not afford to buy very much of them.

Thus, both preferences and purchasing power determine individual demand curves. These demand curves can be drawn and compared in the same diagram, as in Figure 3. Marginal utility curves, in contrast, cannot be directly compared among people. The fact of comparability among demand is important. By contrast, your preferences for goods are private matters, hard to express in words or numbers. Also, you can't really compare your utility levels numerically with those of other people. Yet what you will *pay*, in money, converts those noncomparable attitudes into money values that are precisely comparable and measurable. They are definite and have a common basis.

The individual demand curves will vary from person to person. Some demand curves will be high; others will be low or in the negative range, as Figure 3 illustrates. Some will slope steeply, others gently; some will be straight lines, others curved or wavy. Yet all of them will slope down, reflecting the "law" of diminishing marginal utility.

For each price, each buyer will take a specific amount. For example, in Figure 3, when gasoline is $1.25 per gallon, the amounts purchased by Thompson, Gottschalk, Martinez, Stone, and six others are as shown. Different prices will result in different quantities being purchased by each person. One person will buy no gasoline at all at the going price.

Scarce goods and free goods

Scarce goods Economic goods are scarce goods. They have prices attached to them, set by the interaction of demand and supply. They are scarce because costs have to be incurred to supply them. If the price is not paid, the supply is not provided: Stores will not provide goods to you (such as clothes, gasoline, or groceries) unless you pay the prices.

On a deeper level, suppliers cannot continue producing costly goods unless consumers pay them enough to cover their costs. Consumers are willing to pay a price for the good because it is in the range of

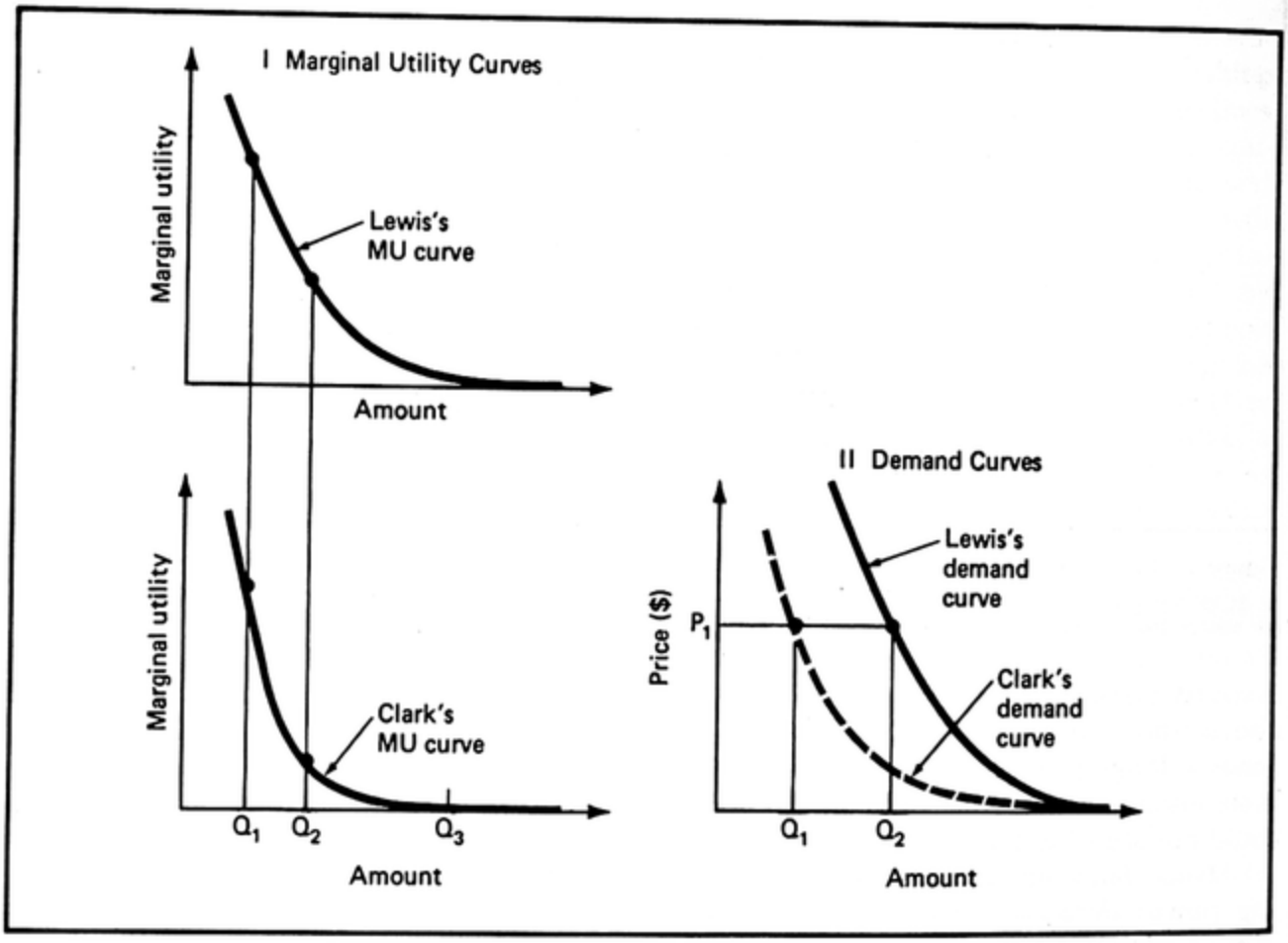


Figure 4 Two people's demand for fresh air: An illustration

Clark's and Lewis's marginal utility curves for fresh air are as shown. Both people would be willing to pay for fresh air when it is scarce, as shown in Panel II. But Lewis's demand curve lies above Clark's, because Lewis either has higher marginal utility, or is richer, or both. Therefore, at a price such as P_1, Lewis will buy more fresh air, for example, by taking more trips to the country, or by living in a more expensive but better ventilated part of town.

scarcity in their marginal utility and demand curves. Since marginal utility is positive, consumers will want more units of the good intensely enough to pay money to get them.

Free goods But many valuable goods are so abundantly available that two conditions hold: (1) the cost of supply is zero, and (2) consumers use them at rates that bring them into the range of zero marginal utility. Such free goods include natural resources like air and sunlight. All people value and consume them in large amounts, but they are free. If they were scarce, then people would have high marginal utility for them and be willing to pay for them.

As illustrated in Figure 4, a limited quantity of fresh air at Q_1 would leave the two people with high marginal utilities. Lewis's and Clark's marginal utility would be less if the quantity of fresh air increased to Q_2, but it would still be positive. By the quantity level Q_3, their marginal utilities are virtually zero.

Their demand curves reflect both their preferences and their purchasing power. Lewis, being richer than Clark and also possibly liking fresh air more intensely, has a higher demand curve than Clark. If

fresh air were priced at P_1, they would buy amounts Q_b and Q_c. Varying prices would cause varying purchases, as shown by the demand curves. Since the price is zero, they both use about Q_{bc}.

Moreover, many "free goods" do involve costs, even though they are provided at a zero price. Two important categories are: (1) many goods provided by governments, and (2) prepaid goods. Governments provide many "public" goods and services at a zero price, even though the goods are costly. Examples are roads, parks, schools, police and courts, and the national defense. Some of these are consumed as a matter of choice, especially roads and parks. ***Because they are provided free, people use them to the point where their marginal utility is zero.*** That may be well above the levels of consumption that would be chosen if a price were imposed.

Prepaid goods are also consumed up to the quantity where marginal utility is zero. The consumption of food is often a good illustration of such prepaid goods. In college dining halls and at the family table, the consumer usually does not pay a specific price for each item chosen. The dining hall bill is paid in advance, and family members are certainly not asked to pay for each thing they eat. Commonly the food is abundant, and often there are seconds available on many items.

Therefore, the consumers' choice is often as shown in Figure 5. The average cost of providing the food may be $1.50 per meal. Yet at the point of choice, the price of the food is zero. Therefore, the person eats the amount *A*, where the demand curve is at zero (the effective price of the meal). That quantity is more than the amount *B*, which would be the amount chosen if each item were paid for separately. Since the food appears to be free, the person consumes it to the point where its marginal utility is zero. The result is often more total utility, more eating, and more weight!

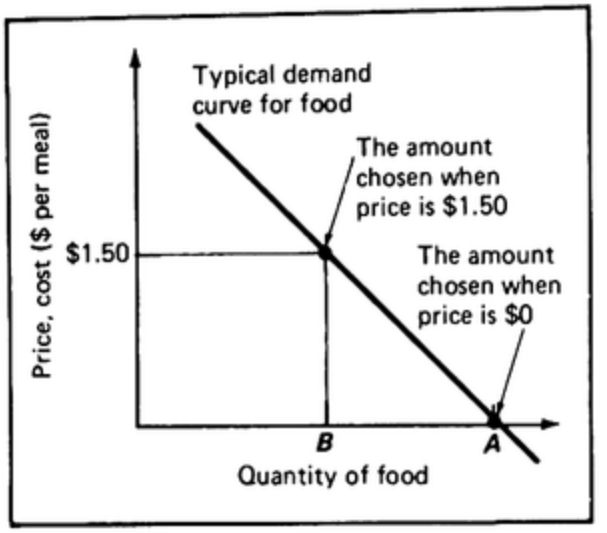


Figure 5 The amount of eating at a dorm may not reflect price

The consumers' demand has a degree of price elasticity. If charged a higher price, the person would eat less. Since the meals are not priced individually and people can take extra amounts at no additional charge, eating will proceed out to Point *A*. If the $1.50 average price were charged and no extras were permitted, only the amount *B* would be eaten. The difference between *A* and *B* is: (1) overeating, (2) healthy nourishment, (3) economically inefficient, or (4) all three?

Would another pricing scheme (such as item-by-item meal tickets for students) be more efficient? That depends on the elasticity of demand for food. If the elasticity is low, then the amount of "extra" eating at *A* is not much above *B*. But if elasticity is high, then students may eat much more because the food is "free." (Ultimately the food is not free, for it must be paid for by the family or in the student's total yearly food bill. It just seems "free" at the point of consumption. That is the special feature of the pricing.)

Marginal utilities and prices in equilibrium

Economists also use marginal utility to define the conditions that the consumer reaches in equilibrium. We now show those equilibrium conditions.

Reaching the equilibrium Each consumer allocates his or her spending among hundreds of items. This allocation is done

intuitively, in line with personal preferences. For example, you may suddenly feel that you have come to like movies more than eating out. Accordingly, you will rearrange your spending toward movies until it fits your new preferences. Your whole purpose is to allocate your spending so that the resulting set of goods will *maximize your utility*. When this best allocation is reached, you will feel that no further changes in spending will increase your satisfaction. This process is done routinely every day by millions of people. They simply do what seems best to them, without preparing any precise formulas or technical details.

But economists can define that process precisely, using the concept of marginal utility. *The best allocation of income on goods will be reached, they say, when the consumer reaches amounts of goods that give this set of equalities:*

$$\frac{\text{Marginal utility of Good 1}}{\text{Price of Good 1}} = \frac{\text{Marginal utility of Good 2}}{\text{Price of Good 2}} =$$

$$\frac{\text{Marginal utility of Good 3}}{\text{Price of Good 3}} = \ldots \frac{\text{Martinal utility of Good } n}{\text{Price of Good } n}.$$

Utility is maximized, whether the consumer does it by intuition or by deliberate estimates of actual utilities. These equalities are the conditions reached by a consumer when utility is at its maximum.

To understand this set of marginal conditions, concentrate first on the meaning of an individual ratio. Marginal utility is the increase in satisfaction that you receive in each period from consuming the final unit of the good. If marginal utility for the last ball-point pen you bought is 10, for example, then that last pen contributed 10 units to your satisfaction. If you paid \$1 for the pen, then marginal utility/price = 10/\$1 = 10. The final dollar that you spent on the fifth pen gave you 10 units of satisfaction or utility. ***The marginal utility/price ratio*** simply shows the amount of utility received from the last dollar spent on the good.

To examine the meaning of an equality among marginal utility/price ratios, economists usually begin with examples involving two goods. Suppose that your present pattern of consumption is seven pens and five magazines per month. The marginal utility of pens = 10, and the marginal utility of magazines = 50, while the price of pens = \$1 and the price of magazines is also \$1. The marginal utility/price for pens = 10/\$1 = 10, while the marginal utility/price for magazines = 50/\$1 = 50. The last dollar that you spent on pens gave you 10 units of satisfaction, while the last dollar you spent on magazines gave you 50 units of satisfaction.

These calculations look impossibly precise, but they show the same result that you would reach intuitively. The ratios are not equal: *You are buying too many pens and too few magazines.* The last dollar that you spent on magazines gave you five times more satisfaction than the last dollar that you spent on pens.

To increase your total satisfaction without spending any more money, simply switch \$1 or more from pens to magazines. When you spend \$1 less on pens, you give up approximately 10 units of satisfaction, but that \$1 now spent on a magazine will gain approximately 40 units of satisfaction (now 50, because of diminishing marginal utility). The net gain in satisfaction or utility would be 30. As you continue shifting dollars and consuming more magazines, the marginal utility of magazines will decrease. You purchase fewer pens (moving back up the marginal utility curve), and so the marginal utility from pens at your spending level will rise. As the switching continues, the marginal utility/price ratio for magazines falls while the

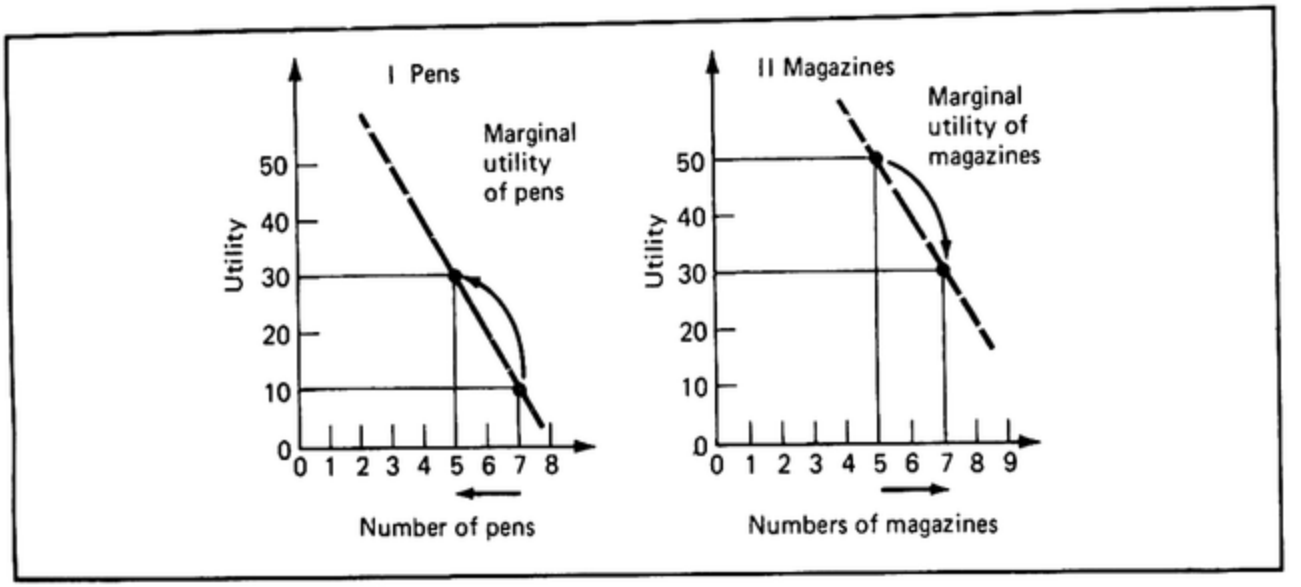


Figure 6 Reaching an equilibrium between two goods

Initially you consume 7 pens and 5 magazines per month. But the MU/price ratios are unequal: 10/$1 = 10 for pens and 50/$1 = 50 for magazines. You cut back on pens, switching the money to buy more magazines. At 5 pens and 7 magazines, the ratios are both 30/$1 = 30, and you are in equilibrium. The adjustment has increased your total utility by 40.

Table 3 Illustrative marginal utility values for pens and magazines

Pens		Magazines	
Number	Marginal Utility	Number	Marginal Utility
1	70	1	90
2	60	2	80
3	50	3	70
4	40	4	60
5	30	5	50
6	20	6	40
7	10	7	30
8	0	8	20

marginal utility/price ratio for pens rises. At some point, the ratios will come to be equal.

The process is illustrated in Figure 6, whose curves are derived from Table 3. The marginal utility curve for pens is shown in Panel I, for magazines in Panel II. While moving back up the MU curve for pens, you free dollars that can be transferred to magazines, which provide higher marginal utility. At 5 pens and 7 magazines, you reach equilibrium, with

$$\frac{\text{MU pens}}{\text{Price pens}} = \frac{30}{\$1} = 30$$

$$= \frac{\text{MU magazines}}{\text{Price magazines}} = \frac{30}{\$1} = 30.$$

By shifting from the first situation, you have lost 30 units of satisfaction from pens (10 plus 20) but gained 70 units from magazines (40 plus 30). Total utility has risen by 40, while total spending on pens plus magazines is unchanged at $12 per month.

The general rule is: Unless the ratios of marginal utilities/price for all goods are equal, you can always reallocate your dollars and increase your total utility, while keeping expenditures constant. Once the marginal utility/price ratios are equal for all goods consumed, then the last dollar spent on each good gives the same amount of satisfaction or utility. *At that point, no further reallocation will increase your satisfaction.* You are receiving the highest total utility that you can receive, given your preferences and the constraints of your budget. Since there is no reward or incen-

tive for additional change, an equilibrium has been reached.

The equilibrium reflects a reconciliation between (1) your *preferences* and (2) the *external limits* set by your income and by the prices of goods. You don't just "buy what you want." Instead, you consider the cost of each good to you (its price), and compare that against the marginal benefits to you (its marginal utility). Here again is the familiar economic comparison of *costs* with *benefits*. Consumers do it repeatedly to reach their individual best allocations.

Restoring an equilibrium Suppose that you have indeed achieved an equilibrium in your consumer spending. The marginal utility/price ratios for all goods that you consume are equal. You feel satisfied with the allocation of your money. Now consider two main changes—in prices and in preferences—that may disturb the equilibrium.

THE PRICE OF ONE GOOD INCREASES For that one good, the marginal utility/price ratio will now be lower than for other goods. To restore equality, you will have to buy less of this good and more of other goods. This decrease in the purchase of the good whose price has increased ties in with the *substitution effect*, which we noted (in Chapter 4) as one reason for the downward slope of the market demand curve. The substitution effect states that as the price of a good increases, alternative or substitute goods always become relatively more attractive. That also fits market realities: As the price of an individual good increases, most consumers will purchase more substitute goods.

By focusing on the marginal utility/price ratios, economics shows why other goods become more attractive as the price of a particular good increases. The resulting adjustments may be negligible for an inexpensive item. But if a large item such as housing or food has a big price change, the resulting shifts in purchases of all goods may be substantial.

YOUR PREFERENCES MAY CHANGE Suppose that your interest in magazines suddenly drops off. Since you get less utility from them, their marginal utility curve from Figure 6 shifts downward, as shown in Figure 7 and Table 4. The MU/price ratio for the seventh magazine is now 10/$1 = 10, definitely unequal to the pens' ratio of 30. You now adjust by cutting back on magazines, moving back up that MU curve. Using the money from magazines, you move down the MU curve for pens. Equilibrium is reached quickly at six each of pens and magazines, where the ratio is 20/$1 = 20 for both. The result also fits common sense: When you lose interest in something, you will buy relatively less of it.

Shifts versus movements along demand curves

As with market demand curves so it is with individuals' demand curves: It is important to distinguish between shifts of demand curves and movements along them. The whole issue of shifts versus move-

Table 4 Adjusting to a drop in utility from magazines

Pens		Magazines	
Number	Marginal Utility	Number	Marginal Utility
1	70	1	70
2	60	2	60
3	50	3	50
4	40	4	40
5	30	5	30
6	20	6	20
7	10	7	10
8	0	8	0

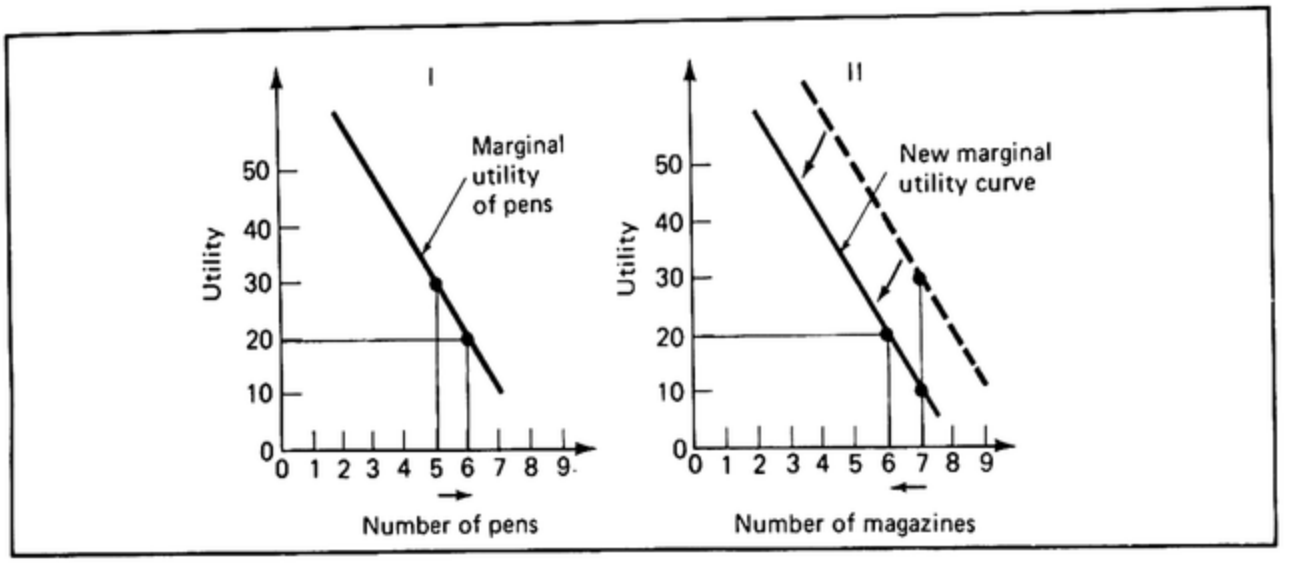


Figure 7 Adjusting to a drop in utility from magazines

The new marginal utility curve for magazines is below the old one. If you still consume 5 pens and 7 magazines, the MU/price ratio for magazines is now only 10/$1 = 10, well below the 30/$1 = 30 ratio for pens. So you switch from magazines to pens as shown, to reach a new equilibrium at 6 pens and 6 magazines and MU/price ratios of 20.

ments along curves was dealt with in Chapter 4. To summarize that discussion: **Quantity demanded refers to a particular point on a demand curve, while demand refers to the entire demand curve. "A change in quantity demanded" refers to a movement along the demand curve from one price-quantity combination to another. *Such a movement can only be caused by a change in price.***

Individual demand has great power to explain human choice. Think of all consumers as making these balancing choices and adjustments among everything that they buy. ***People are all marginalists in their consumption choices, making and adjusting such selections day in and day out.*** **In fact, almost every decision you make involves such a balancing among its *benefits* (marginal utility) and its *costs* (the price you have to pay) *at the margin.***

Few people think about such choices in the precise economic terms of equating the ratios between marginal utilities and price. Yet, they usually behave *as if* they were doing roughly that: adjusting the levels of activity so as to bring marginal benefits (utilities) and costs into line.

Consumer surplus

The demand curve shows an important condition of consumer demand: *Consumers receive more value from purchasing a good than the money value they pay to the supplier. This extra value is termed consumer surplus.* It is shown quite simply with a standard demand curve, as in Panel I of Figure 8.

First note that you pay $20 to buy 20 units at $1 each. That $20 is shown by the area of the unshaded rectangle Prices times Quantity ($1 times Q_1). The $20 is your sacrifice to get all 20 units. You chose to take 20 units because the 20th unit is just barely worth its $1 price to you.

But you receive the other 19 units, all of which you were willing to pay more than $1 for. Your demand curve shows that you would have paid $5 for the 5th

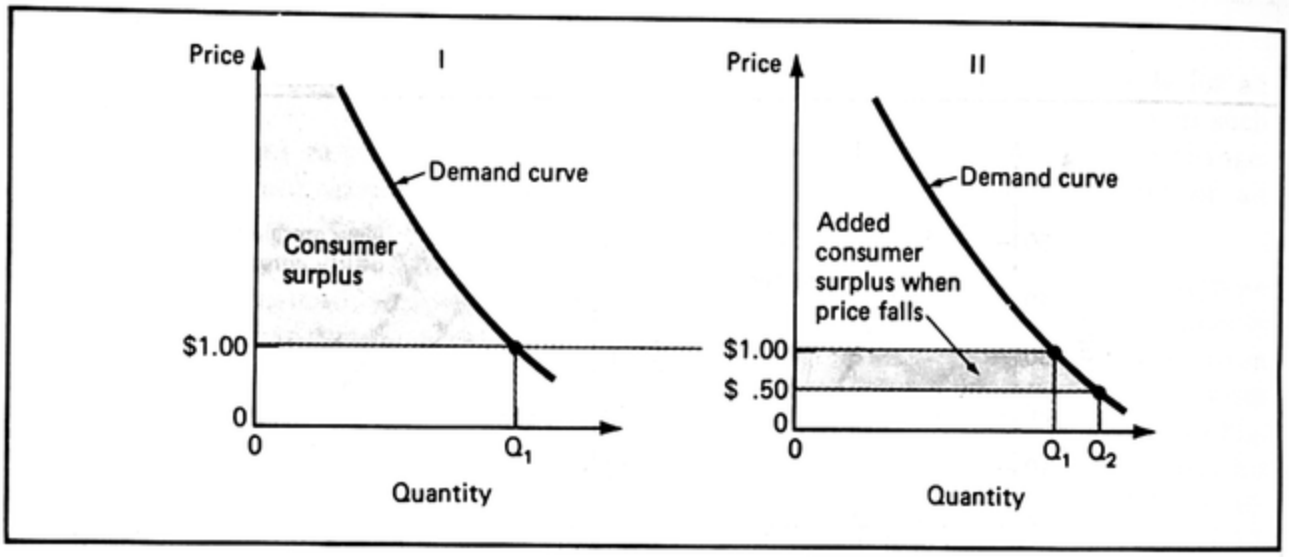


Figure 8 Consumer surplus

When price is $1.00 and the consumer buys Q' units, then ($1.00) Q' is paid for the good. The consumer also receives the benefit shown by the shaded area under the demand curve in Panel I. This is the consumer surplus. If the price falls to 50 cents, as in Panel II, an added amount of consumer surplus is realized, as shown by the shaded area between $1.00 and 50 cents.

unit and $2 for the 14th unit, if it were necessary. Since the price is $1, you only have to pay $1 each for all 20 units. Therefore, you receive a surplus value above $1 on 19 units. That value is shown by the area under the demand curve and over the $1 price line.

This consumer surplus is a universal phenomenon, occurring whenever a consumer buys more than one unit of a good. By establishing one price for the good, the market assures that consumers can gain extra value above what they must pay for the good. The lower the price, the larger the consumer surplus is. If the price should fall to 50 cents, for example, as in Panel II of Figure 8, then all of the shaded area is added to the consumer surplus. If the price dropped to zero, making the good a "free good," then all of the area under the demand curve would be realized as consumer surplus. (There are many such cases. Public goods such as roads, public schools, fire protection, and many parks are provided at no price to the individual user.)

The amount of consumer surplus for each good depends on the elasticity of the consumer's demand curve, as shown in Figure 9. A high elasticity of demand means that consumer surplus is small (Panel I), while a low elasticity goes with a large consumer surplus (Panel II). That is entirely logical. Inelasticity exists for goods that are urgently wanted, such as a life-giving drug or other such "necessities" as water, food, and housing. *When urgent wants can be met at a low price, you receive a large consumer surplus.* Highly elastic goods, by contrast, are those you want only mildly. You would give them up entirely if the price rises a little.

Consumer surplus is an important economic phenomenon. Though the consumer decides by focusing on the marginal units, the result is extra value on all the nonmarginal units of the good. Economists use the concept of consumer surplus frequently in evaluating the efficiency of the economy.

Market demand is the sum of individual demands

Deriving market demand from individual demand is a short and easy step. Simply add up the individual curves horizontally to get the market demand curve. The market, after all, is simply the sum of the actions of all the people acting in it.

The process of summation is illustrated in Figure 10, using the ten individ-

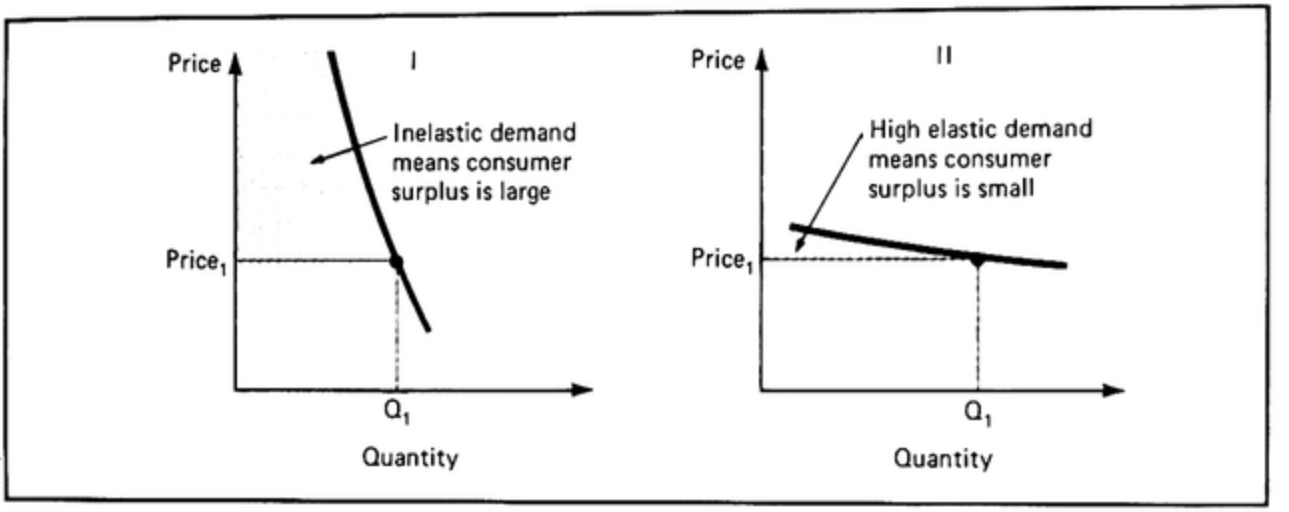


Figure 9 The amount of consumer surplus depends on elasticity of demand

ual demand curves shown in Figure 3. For the moment, assume that the market contains only these 10 consumers (many markets have millions of consumers, while others have as few as 50 or less). At a given price, each consumer in the market will take a definite quantity, which may be zero or larger. For example, at a $1.50 price, the ten consumers whose demand curves are shown in Panel I of Figure 10 will together buy exactly 150 gallons per day (with one of them buying none). If the price were $3.00, these ten consumers would purchase only 52 gallons per day (with six of them buying none). At a price of $3.50, only one consumer would purchase any gasoline at all, approximately 13 gallons per day.

As the quantity demanded by each consumer at each price is added to the others, the market demand curve is given precisely. The curve shows the alternatives that consumers would take at different prices. At any given time the market out-

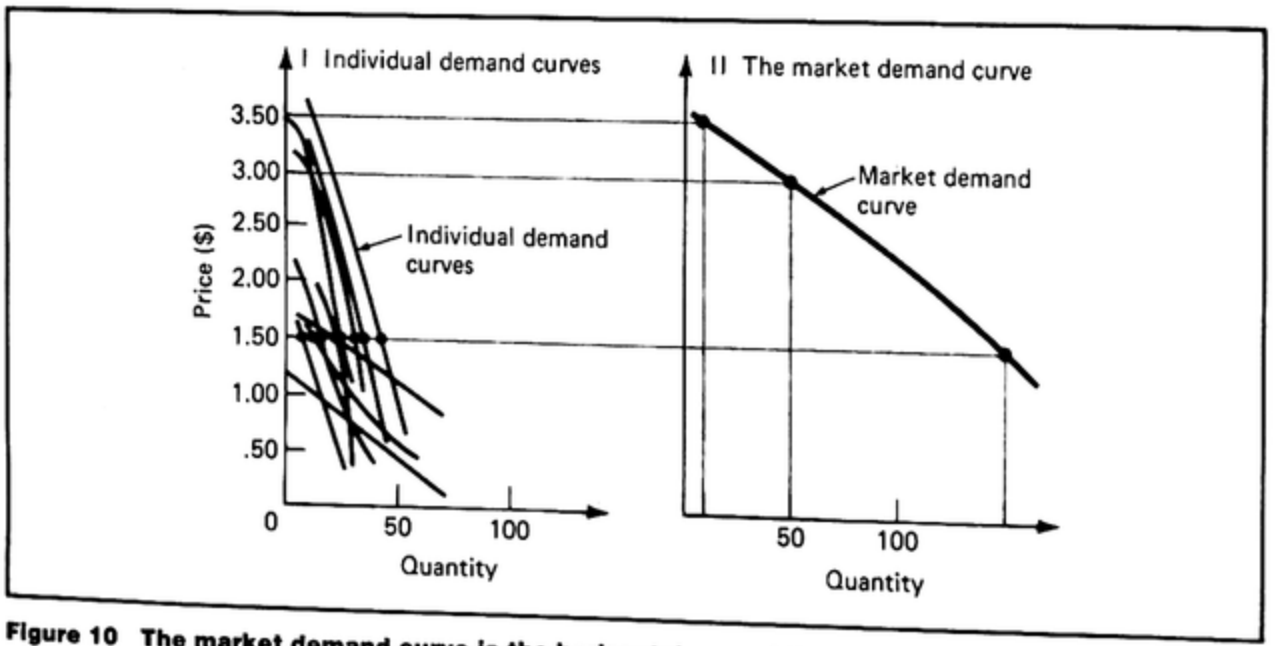


Figure 10 The market demand curve is the horizontal sum of the individual demand curves

At $1.50 per gallon, each person takes the amounts as shown in Panel I. They, plus the other consumers in the market, take 150 gallons per day altogether, as shown in Panel II. The amounts at other prices are traced out by the same horizontal addition.

come is just one point on the demand curve, which shows the actual quantity that was bought at the given price. The complete demand curve merely shows what might have happened at alternative prices.

Consumer surplus also exists on a summed basis under the market demand curve. The analysis is the same as for individual demand. At the going price, consumers pay an amount shown by price times quantity. They also receive consumers' surplus, as shown by the area under the demand curve but above the price. If price falls, the area of consumer surplus increases, parallel to the lesson of Figure 8, Panel II. The size of the consumer surplus depends inversely on the elasticities of the market demand curve, similarly to the patterns in Figure 9. If demand is inelastic, the consumer surplus is large; highly elastic demand provides only a small consumer surplus.

Derived demand

So far we have been discussing only the demand of households for goods that go directly for household use. This is often called "final" demand, for it comes at the end of the chain of production. But that chain—which goes from raw materials, to processed materials and parts, to semifinished goods, and to final goods—can have many stages. At each stage, demand and supply interact to set prices and quantities.

The demand at each early stage is called "derived demand," for it is derived backward from the final demand of households. For example, households buy bread. To bake the bread, bakers have to buy flour. Their derived demand for those goods is met by flour companies. The flour companies, in turn, buy grain; their derived demand goes back to the farmers who grow the grain. The farmers, in turn, have a derived demand for tractors, seeds, gasoline, fertilizers, and so on.

Derived demand arises in companies, not households, but the analysis for it is much the same as for households' final demand. Each firm has its own individual demand curve, and the market demand curve is the sum of these individual curves. Like personal demand curves, these derived demand curves can shift if other variables change. There can also be substitutes and complements in derived demand, just as in final demand.

The validity of demand analysis

The analysis of individual demand is crucial to microeconomics. Here, as in other parts of economics, we focus on marginal choices, which weigh the costs and benefits of additional units. Yet, doubts about the microeconomic theory of demand have been vigorously debated ever since marginal utility analysis emerged in the 1870s. Since some of them may have occurred to you, we pose them here. The answers will show that the analysis of demand has cogency and power in explaining the mass of consumer behavior.

Question: Aren't these measurements and ratios really *too precise to believe?* No consumer could possibly hope to make choices with such perfect exactitude.

Answer: People *tend* toward rational choices, *as if* their decisions were guided toward equal marginal utility/price ratios. They may indeed never reach the desired equality conditions, for life is full of changes and approximations. But even if consumers are always out of equilibrium, making a series of rough-and-ready choices, the ratios are still valid for explaining the main directions of their choices.

Question: Aren't the marginalist concepts *irrelevant to real choices?* Nobody actually uses these concepts when making real choices about what to buy. Even the most zealous marginalist-trained economist doesn't calculate marginal utility/price ratios when buying a loaf of bread, or a stereo, or any other good. Instead, people make rough choices one by one, using hunches, impulses, or feelings.

Answer: No, the concepts are still valid. As long as people are trying to obtain the greatest satisfaction they can, within the constraints on their purchasing power, their marginal decisions will be much the same as if they had gone through the analysis and made precise measurements. There will often be mistakes and imprecision, but the whole outcome will be in line with the analysis.

Question: *The consumer is treated as selfish,* interested only in maximizing his or her own satisfaction or pleasure. Doesn't this ignore the altruistic or charitable actions of many consumers?

Answer: The analysis does focus on people's buying choices, which are mostly based on self-interest. Charity and help to others do not fit neatly into the economic calculations. Such unselfish action can be included, though, by recognizing that they give pleasure or satisfaction to the giver as well as to the receiver. In other words, $10 given to a charity may yield the giver the same amount of pleasure or utility as $10 spent on a bottle of wine. Indeed, the analysis would say that the marginal utilities of the two $10 payments *will* be equal.

Question: The pleasures from consuming *economic goods are only one source of happiness.* There are at least two other major sources. First, the best things in life may be free, or at least many of them are. Love, health, a beautiful day, good friends—these and many other important things in life can't be bought with money. Indeed, paying cash for them would often spoil them. Second, a person's job is often a prime source of meaning and satisfaction. After all, it occupies most of one's waking hours and thoughts. Success at work often dwarfs the satisfaction from anything that is bought. For these two reasons, aren't consumption choices only a sideshow to the real sources of happiness?

Answer: This can be true, especially for people who are young, well off, and in fine jobs. Even so, the consumption choices are important. Moreover, for the rest of the population—the majority who are not so well favored—the spending choices are more important. For older people with mediocre jobs, for example, what money they have must be spent carefully to avoid serious troubles: debts, family quarrels, skimping on necessities, loss of status, and so on. For most people, consumption choices are urgent, and making rational choices can improve their whole sense of well-being. As for job choices, marginal utility is, in fact, the basis for explaining them, as we will show in the chapter on labor economics.

Summary

1. A consumer's goal is to *maximize utility,* that is, to select the array of goods that will result in the greatest amount of utility or satisfaction, within the constraints of the consumer's budget.
2. As more of a good is consumed, total utility or satisfaction rises, at least up

to a point. Yet each additional unit of the good contributes less and less to total utility. This addition to total utility from consuming an additional unit of the good is called *marginal utility*.

3. The downward slope of the individual demand curve expresses the *law of diminishing marginal utility:* Marginal utility declines as additional units of the good are consumed.
4. The shape and height of the marginal utility curve, expressing *willingness* to buy a particular good, will influence the shape and height of a consumer's demand curve for that good. The height of a consumer's demand curve for a particular good will also be influenced by income—the consumer's ability to purchase the good.
5. When a good's price is zero, it is a free good. Such a good will be consumed until its marginal utility is zero.
6. When a consumer has reached an equilibrium or balance in consumption, the marginal utility/price ratios for all goods consumed will be equal.
7. Distinguishing between changes in quantity demanded (movements along the demand curve) and changes in demand (shifts of the demand curve) is just as important in dealing with individual demand curves as in dealing with market demand curves.
8. Consumer surplus is the difference between the total value of the good to the consumer and the money value that has to be paid for it. It is the area under the demand curve but above the price of the good.
9. To derive the market demand curve for a good, simply add up the individual demand curves horizontally.
10. Although utility cannot really be measured or calculated in practice, consumers behave in about the same patterns and intuitively make the same choices *as if* they were actually measuring utilities and equating marginal utility/price ratios.

Key concepts

Rational choices to maximize utility
Utility
Law of diminishing marginal utility
Total utility
Marginal utility
Marginal utility/price ratios

Questions for review

1. Think about the range of goods available for you to consume.
   a. Can you think of any goods whose total utility for you must be negative for every unit? What is the clue that these goods have a negative total utility for you?
   b. Consider some of the goods that you actually consume each week. Can you make any intuitive guesses about which goods have:
      i. sharply declining marginal utility curves
      ii. fairly flat marginal utility curves.
2. Suppose that, for two goods you consume:

$$\frac{MU \text{ pickles}}{P \text{ pickles}} = \frac{MU \text{ ice cream}}{P \text{ ice cream}}$$

   a. What would this equality mean? Explain carefully.
   b. Suppose that the price of pickles now increases.
      i. What happens to the equality?
      ii. What kind of adjustment would you make in your purchases? Explain.

7

The Enterprise

As you read and study this chapter, you will learn:

- the broad patterns of private business enterprises in the U.S. economy
- other types of enterprises, and the main kinds of industries
- the basic nature and choices of enterprises
- the main indicators of success for the firm
- typical conditions in the creation and growth of a new enterprise

Now we take you across the great divide in microeconomics, moving from the demand side and its households over to the supply side and its enterprises. You already know that the supply curve slopes upward, has elasticities, and helps to determine the market equilibrium. We will now present the foundations of the supply curve, showing how the costs of firms in each competitive market add up to form the supply curve.

In analyzing firms, economists concentrate on certain basic, clinical concepts. Yet actual businesses involve astonishing color and variety, with much human drama and strife. Moreover, the supply side is pervaded by deep social issues. Each company is a workplace, an expression of the people who work there year after year. The pale word "firm" covers all manner of tiny, giant, robust, and decrepit enterprises in a variety of industries.

The economic analysis of supply is about the reality of business activity, not some remote abstraction. Enterprises provide jobs and careers, establish prices and outputs, develop technology, make family fortunes, and set much of the social and cultural tone of society. Most of your relatives probably work for business firms, and perhaps you will too in due course. You already know many enterprises and industries, some of them intimately, because you deal with them every day.

This chapter presents some of that reality, setting the stage for the concepts of production and cost that follow in the next two chapters. First we give the main lines of the business population in the range of real industries. We also cover conglomerate firms, stock ownerships, and *non*private forms of enterprises.

In the second section of this chapter, we survey the main practical kinds of inputs, outputs, and production conditions that firms have. We then show how firms organize differently to produce single products and a wide diversity of products. We also present the main success indicators of firms (profits and stock prices).

Finally, we draw the parts together in a case study, tracing a small firm through its first few years. Throughout we aim to prepare you to understand cost and supply both in the abstract and as they can be seen directly in real enterprises.

Patterns of actual enterprises

Some patterns of actual enterprises are familiar to many readers. Other features are less well known. We present private firms first, for they are the most common form in the United States and other Western economies. Next come other types of firms, such as public enterprises and cooperatives. Then we note the variety of industries within which firms operate.

Private enterprises

A ***private enterprise*** is a firm owned by individuals or by other firms and operated with the primary aim of making profits for its owners. There are almost 14 million such private firms in the U.S. economy, ranging from tiny corner shops to the largest corporations. They all seek to ***maximize profits,*** even though their specific settings and choices vary enormously.

The private business population in the United States includes a range of firms from small to large. Small businesses are the most numerous by far: 13.4 million out of the total 13.7 million firms sell less than $500,000 worth of goods each per year. However, since those small firms together account for only 17.1 percent of total sales, many observers regard the large firm—especially the large corporation—as dominant in the U.S. economy. Yet, small and medium-sized firms are very important and worth considering in some detail.

Small business The smallest businesses include about 11 million units that take in less than $100,000 in revenue each year. These are the familiar neighborhood stores and tiny workplaces that are run by one or two people, with a small, local clientele. Some of these shops have existed for decades, but most of them close after a short time.

Most owners worked at a production job before starting their own business with a small amount of capital, and often a dedicated spouse. Usually the space and equipment are rented rather than bought. Joining the legions of small business is remarkably easy in this free economy.

What's hard is making a profit. We will trace a case study of a successful small

business in the third section of this chapter, showing the severe stresses such a business entails. Unless the firm has a large special advantage, it must contend not only with the internal problems of getting started but also with competition from other firms. Production must be organized, and customers must be attracted who will pay enough in revenues to cover the costs of the business. It may be necessary to offer special low prices at first to draw in new customers. The firm must establish an identity, as well as produce efficiently so that its costs per unit are below the prevailing market price. And it takes time to set up production, train people, get secure supplies of inputs, and develop demand for the firm's product by advertising and other methods.

Accordingly, most new businesses disappear within two years. Some owners go bankrupt, but most simply close down, pay their debts, and return to other jobs, wiser, sadder, and poorer. Successful firms, however, may sell out at a handsome gain or merely continue as a rewarding way of life for their owner-managers.

Most small firms are family held and do not become corporations. But many small firms do incorporate, and nearly all medium-sized and large firms are corporations. We now must consider corporations and their role.

Corporations ***Corporations are firms that issue voting stock, which investors can buy and sell.*** These owners (the stockholders) are not responsible for all of the corporation's debts or actions. The most they can lose financially is the value of their shares, if the price of those shares should fall to zero in the stock market.

The corporation is the dominant form of U.S. business, and has been for many decades. That is true also of West European industrial economies and of Japan. But the domination of large corporate business has distinct limits.

That can be seen by considering medium-sized firms. There are about 35,000 U.S. firms with yearly sales revenues between $10 million and $1 billion. They can be regarded as medium-sized, compared to the approximately 425 large firms that have sales above $1 billion.

These medium-sized firms are usually substantial, well-established companies, producing a variety of products and selling on a regional or national scale. They have continuity and identity in the market. Their management is usually specialized to handle complex tasks with professional skill. Of course, many of these firms encounter special problems, and some of them eventually fail. But as a group, these firms' quality, size, and continuity distinguish them from the mass of small businesses.

These 35,000 medium-sized firms form the bulk of the U.S. economy. Their combined sales and assets are over half of the economy-wide totals. In all sectors except farming and services (two sectors where the small family firm is dominant), these medium-sized firms are a major factor.

Large corporations The largest corporations include about 300 manufacturing firms with sales above $1 billion, plus about another 125 in such other sectors as banking, retailing, transportation, and utilities. These firms operate on a national or global scale, often in many different industries. They have at least 15,000 workers each, with the largest seven of them employing over 300,000 workers apiece. The 500 largest firms (of all kinds) in 1981 had about $2 trillion in sales, which was about 40 percent of all sales in the economy.

Portions of the manufacturing, utility, and financial sectors lend themselves to large-scale technology, which, in turn, re-

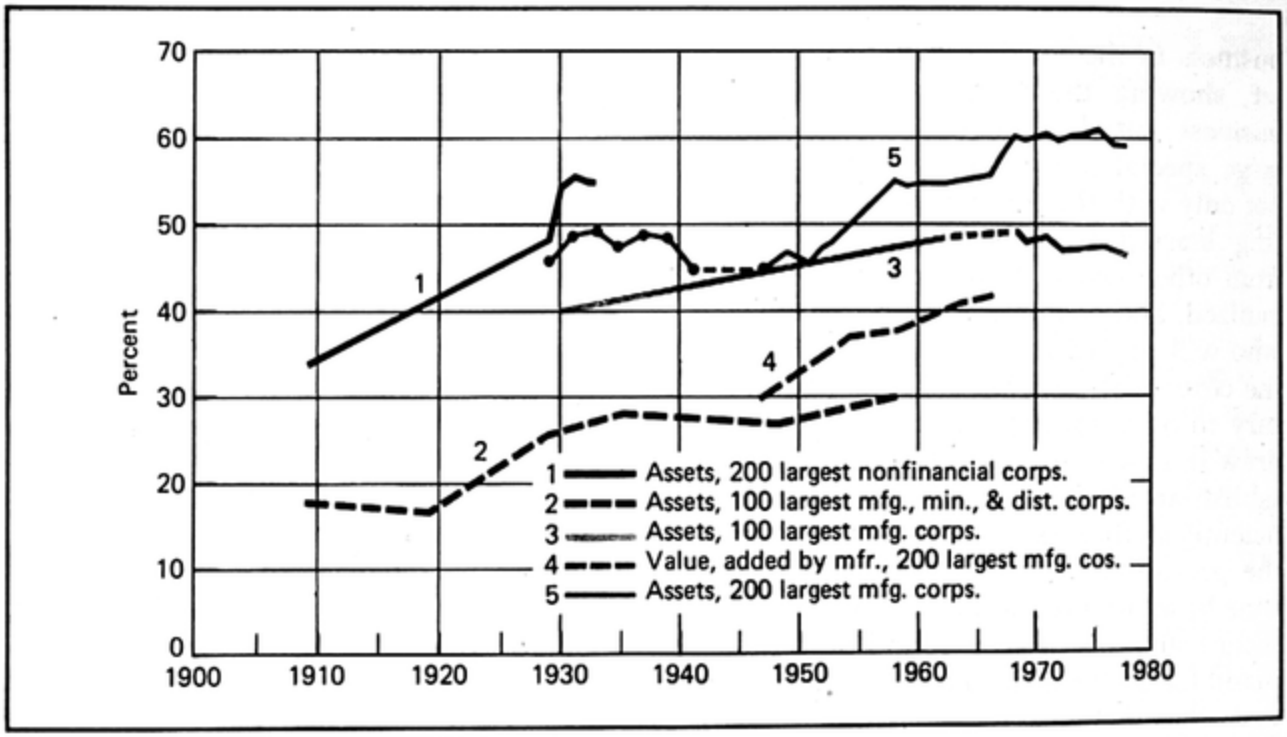


Figure 1 Scope and trends of large U.S. corporations

No single measure of size—assets, sales, employees, etc.—is the sole index of the share of large firms in the economy. Here several alternative measures convey a consistent impression: a strong rise to 1960, tapering off to a possible decline after 1970.

Sources: Adapted from John M. Blair, *Economic Concentration* (New York: Harcourt, Brace, Jovanovich, 1972), Chap. 4; and David W. Penn, "Aggregate Concentration: A Statistical Note," *Antitrust Bulletin* 21 (Spring 1976):91–98.

quires large firms. That is why large corporations are prominent in these sectors.

The role of large corporations in the manufacturing sector is suggested by Figure 1. From 1910 to 1960, the largest 100 or 200 firms significantly increased their shares of total assets and value-added. But their shares appear first to have stabilized in the 1960s and then to have slightly declined in the late 1970s. Therefore, the twentieth-century dominance of the economy by big business now appears to have tapered off and is, perhaps starting to be reversed.

The largest firms The very largest corporations warrant a closer inspection, for they are influential in their markets and in many regions where they operate. These corporate giants are shown in Table 1. They include many familiar names because their products are widely advertised and sold. They also figure prominently in U.S. economic conditions, such as employment, profits, and growth.

Diversification Most of the large firms focus on just one or a few product lines, such as Exxon in oil, IBM in computers, and AT&T in telecommunications. Other firms, however, are highly diversified, with operations in many kinds of products. Such firms—ITT and United Technologies, for example—are often called "conglomerate" enterprises.

The diversification among the largest manufacturing firms is enormous. In 1965 one third of the firms operated in 5 or fewer of the 440 industries defined by the U.S. Census Bureau. Among the rest, diversification was extensive. Some 284 were in 6 to 25 industries apiece. Thirty-three

firms operated in 25 to 50 industries. At the extreme, five firms operated in more than 50 industries each. Some of these firms had more than 100 separate companies—each with its own president and other officers—within their conglomerate structure.

In short, large firms are often highly diversified and have to make complicated choices involving large numbers of inputs and outputs. These choices can be analyzed (as the next chapter will show) with much the same basic tools that apply to simple decisions by single-product firms.

Ownership and control In the millions of small businesses, ownership and control are combined in one person, who makes the decisions and also benefits from whatever financial success the firm achieves. If profits are high, the owner-manager may be able to sell the business at a large capital gain. If there are losses, the owner-manager may lose all of the capital as the firm becomes worthless. Even if the small firm is a corporation, the manager often owns most of the stock, so that ownership and control are still closely combined.

But as size increases, *ownership tends to become divorced from control.* The stocks are bought and sold among many investors, while the managers become a more specialized group who draw salaries and may own little of their company's stock.

This divorce of ownership from control evolved after 1890 as large corporations grew and stockholding became diffused. In a landmark book published in 1932,* Berle and Means argued that this divorce—a "managerial revolution"—had changed the nature of large corporations. The managers were now free from close control and able to run the firms largely as they wished. Since 1932, the trend has continued, so that in virtually all of the largest 1,000 corporations there is no major controlling block of shares.

The board of directors still supervises the executives and has to approve all major decisions. But both the board and the executives are largely independent of stockholder control and can select their own members and set their own guidelines. Indeed, on most boards of directors, the executives themselves hold key positions and dominate the discussions. Single owners or large financial institutions (banks, insurance firms) may hold 2, 5, or even 10 percent of the stock in some of these companies. But control is still largely held by the managers.

This divorce between owning and controlling need not be economically harmful. Instead, it encourages executive continuity and professionalism by replacing the old-style industrial buccaneer with the cool modern manager. This may cause two differences in the manager-controlled corporation. First, actions are usually more predictable and objective, rather than reflecting the personality and whims of a single powerful owner. Second, the managers may focus less on maximizing profits for the owners and more on growth, managerial perquisites; and other results that enhance their own importance.

This second effect, which would dilute the central role of profit maximizing, has been studied closely by economists. So far, they have found only slight hints that it occurs normally. A few manager-controlled firms occasionally depart visibly from profit maximizing. But on the whole, the managerial revolution has scarcely affected the primacy of profit as the central goal of private firms, both large and small.

Stock ownership In recent years, between 20 and 30 million people in the United States have owned stock. Most of these

*A. A. Berle and Gardiner C. Means, *The Modern Corporation and Private Property* (New York: Macmillan, 1932; rev. ed., Harcourt, Brace & World, 1968).

Table 1 ***A selection of the largest firms in various sectors (as ranked by sales)***

Industrial		Main Dimensions in 1980			Average Rate of Profit as a % of Equity, 1976–1980
Name of Company	Main Products	Sales ($mil.)	Assets ($mil.)	Profit ($mil.)	
General Motors	Cars, trucks, buses	57,728	34,581	(−763)	15.3
Ford Motor	Cars, trucks	37,085	24,347	(−1,543)	12.2
International Business Machines (IBM)	Computers, office equipment	26,213	26,703	3,562	21.0
General Electric	Electrical equipment	24,959	18,511	1,514	18.5
International Telephone and Telegraph (ITT)	Telephone equipment	18,529	15,417	894	10.9
E. I. du Pont de Nemours	Chemicals	13,652	9,560	716	14.2
U.S. Steel	Steel and products	12,492	11,747	504	5.0
United Technologies	Aircraft engines, elevators	12,324	7,326	393	13.3
Western Electric	Telephone equipment	12,032	8,047	693	13.7
Proctor & Gamble	Toiletries, household products	10,772	6,553	642	17.6
Dow Chemical	Chemicals	10,626	11,538	805	18.9
Union Carbide	Chemicals	9,994	9,659	890	13.8
Eastman Kodak	Cameras, film, copiers	9,734	8,754	1,153	17.5
Boeing	Aircraft	9,426	5,931	600	19.9
Dart & Kraft	Food products	9,411	4,650	383	14.2
Chrysler Corporation	Cars	9,225	6,617	(−1,710)	4.1
Caterpillar Tractor	Earthmoving equipment	8,597	6,098	564	18.2
Westinghouse Electric	Electrical equipment	8,514	6,812	402	9.4
R. J. Reynolds Industries	Tobacco products	8,449	7,355	670	17.8
Oil Firms					
		Sales ($mil.)	Assets ($mil.)	Profits ($mil.)	
Exxon		103,142	56,576	5,650	16.3
Mobil		59,510	32,705	3,272	16.2
Texaco		51,195	26,430	2,642	13.2
Standard Oil of California		40,479	22,162	2,401	16.0
Gulf		26,483	18,638	1,407	12.4
Standard Oil of Indiana		26,133	20,167	1,915	16.6

Utilities

	Main Products	Operating Revenues ($mil.)	Assets ($mil.)	Profits ($mil.)	
American Telephone & Telegraph (AT&T)	Telephone service	50,791	125,450	6,079	11.7
General Telephone & Electronics	Telephone service	9,978	19,720	477	13.3
Southern Company	Electricity	3,763	11,466	344	9.9
Pacific Gas & Electric	Electricity, gas	5,258	11,295	524	10.5
American Electric Power	Electricity	3,756	10,952	348	10.9
Consolidated Edison	Electricity, gas	3,947	7,459	334	10.1

Retail

	Main Products	Sales ($mil.)	Assets ($mil.)	Profits ($mil.)	
Sears, Roebuck	Household merchandise	25,194	28,053	606	11.3
Safeway Stores	Groceries	15,102	3,338	119	13.1
K-Mart	Household merchandise	14,204	6,102.5	260	16.4
J. C. Penney	Household merchandise	11,353	5,863	233	11.1
Kroger	Groceries	10,316	1,997	94	13.3
F. W. Woolworth	Household merchandise	7,218	3,171	160	10.9
Great Atlantic & Pacific Tea Co.	Groceries	6,684	1,230	—	1.4

Banks (ranked by loans)

	Loans ($mil.)	Assets ($mil.)	Profits ($mil.)	
Citicorp	69,915	114,920	499	14.1
Bank of America Corporation	62,482	111,617	643	15.8
Chase Manhattan Corporation	46,506	76,189	354	10.6
Manufacturers Hanover Corporation	30,348	55,522	228	13.1
J. P. Morgan & Company	25,972	51,990	341	14.5

Insurance (ranked by assets)

	Assets ($mil.)	Premium & Annuity Income ($mil.)	Investment Income ($mil.)	Profits ($mil.)
Prudential	59,778	8,668	3,562	444
Federal National Mortgage Association (FNMA)	58,470	5,203	(N.A.)*	14
Metropolitan	48,309	6,010	3,449	479
Aetna Life	22,270	5,412	1,392	185

Source: Fortune Magazine, *Directory of the 500 Largest Industrial Corporations* and *Directory of the Largest Non-industrial Companies*, annual
*N.A.: Not applicable to FNMA.

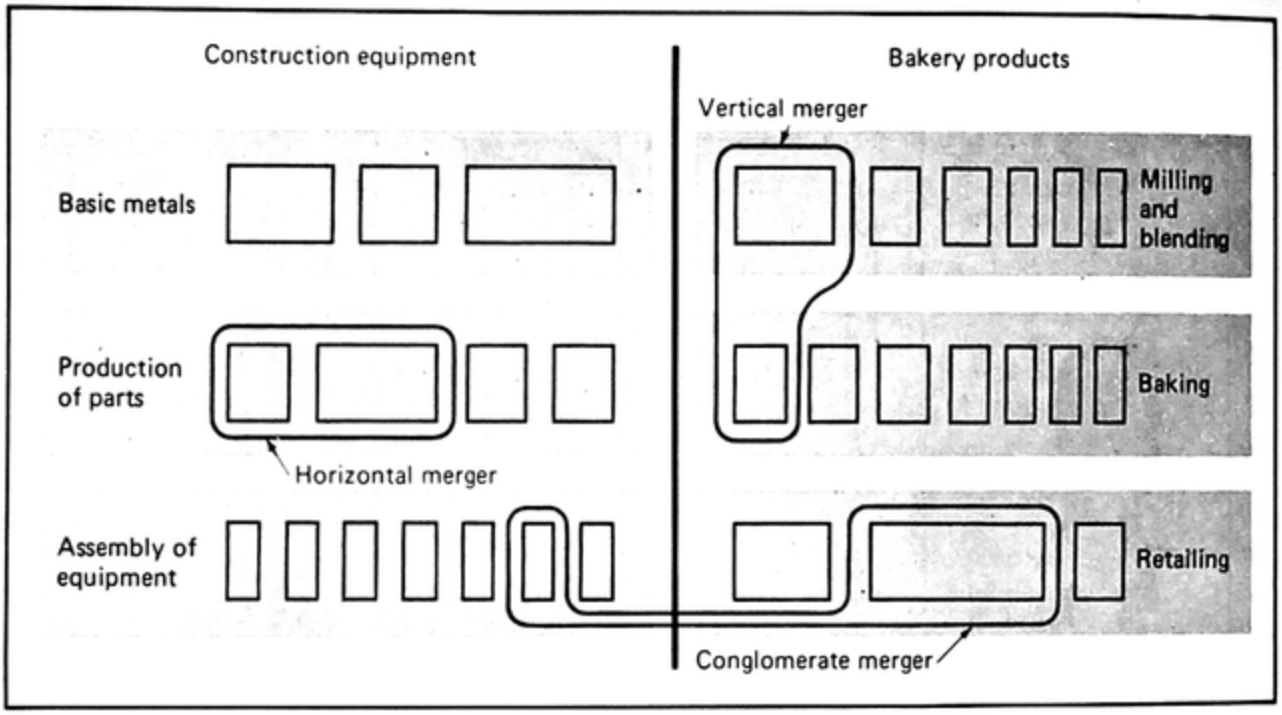


Figure 2 The three kinds of mergers

owners are small investors, with only a few shares. At the other extreme are the wealthiest plutocrats, some with hundreds of millions of dollars' worth of stock. There are also many large "institutional investors," such as insurance firms, pension funds, and banks' trust departments. About three fourths of the trading on the stock exchanges, in fact, is done by these institutions.

The skills of investors vary. The typical small investor is an amateur who learns about stocks mainly from the local newspaper or a local stockbroker. Such an investor usually buys a few shares to hold on to, hoping that their price will rise as the years pass. Only if the price does rise will this person be able to make a (small) gain.

The big-block traders, on the other hand, bring professional skill and detailed knowledge to their operations. They follow conditions minute by minute and their information is thorough. They deal in the most complex, esoteric stock devices. They trade quickly and repeatedly, often acting days or months before crucial information becomes available to small investors (by then, it is too late to be profitable). These professionals can routinely make money on a falling or fluctuating market by using opportunities that the small investor is scarcely aware of.

Mergers ***A merger joins two or more separate firms into a combined firm.*** **Each former firm is now part of a larger enterprise, but without the creation of new production capacity.**

Mergers are numerous and commonplace in the U.S. economy. There is an active market for corporate control: the buying and selling of whole companies. In recent years, there has been an average of about 1,000 mergers a year. Most mergers are small, but some are enormous. The assets acquired totaled about $12 billion a year in the 1970s, but the merger boom of 1980–1982 raised that level to over $18 billion a year.

There are three main kinds of mergers. ***Horizontal mergers*** occur between firms in the same market, as Figure 2 shows. ***Vertical mergers*** join firms at different but related steps in the production chain. ***Conglomerate mergers*** include all the rest, in

A Tour Through *The Wall Street Journal*

The daily facts of the business world are collected compactly each working day in the *Wall Street Journal*. We now take you on a brief tour through the paper, to show where large amounts of information about supply can be found. Try to have a copy of the *Journal* at hand when you read this box.

Individual Companies and Industries

The first 20 pages or so have news stories about companies and industries. Some stories report small events. Others dig deep into the basic problems of companies. Any important event or condition is covered.

Stock and Bond Prices

The last six pages are packed with the stock and bond prices of the previous day. About 5,000 of the country's largest firms are covered in precise detail, as shown in Table A. One finds the price of the stock, its rate of dividends, the number of shares traded, and the range of the stock's price both during the previous day and for the year to date. Similar data are given in the bond tables.

Commodities Prices

Just before the bond tables come tables (B and C) with the prices of about 30 major commodities during the previous day. The products range from pork bellies and coffee to heating oil, plywood, and zinc. Precious metals prices (gold, silver, and platinum) are shown; there are not only current "spot" prices, but also "futures" prices for metal to be delivered 1, 2, 3, 12, and more months in the future.

Interest Rates

There is a small table (shown here as Table D) with the current interest rates for various kinds of loans or bonds. Notice that the rates differ, often for reasons that even the experts cannot explain. Yet, the rates all move broadly together as credit conditions change. In the bond price tables, one can also learn the interest rates paid by specific bonds.

Foreign Currencies

Also nearby is a table (Table E) with the prices of the world's major currencies. These exchange rates are expressed in dollars. For example, a $1.83 price for the British pound sterling means that each pound can be bought for $1.83. There are also futures prices for currencies, showing the present prices for currencies to be delivered in one, two, three, and more months.

Others

One also finds occasional compilations of other economic facts. Regular sections have articles about current conditions in bond markets, stock markets, commodity markets, and foreign currency markets. U.S. automobile sales are reported model by model once a week. Data on world oil production and stocks, foreign stock markets, and national levels of income, money supply, and production are also regularly listed. The first three pages of the second section of the paper give stories about foreign companies and markets.

These and still other data add up to a treasure trove of detailed, precise information. Each day the price activity of

Table A

52 Weeks High	Low	Stock		Yld Div.	P-E %	Ratio	Sales 100s	High	low	Close	Net Chg.
5	2⅜	APlan				3	11	2⅜	2½	2½	.....
17	8⅜	APrecs	s	.32	3.2	6	20	11	10	10	−1¼
3⅜	2⅛	Ampec	n				26	2¾	2⅜	2¾	.....
10⅜	4⅜	ASciE		.35t	7.8	26	60	4⅞	d 4½	4½	− ¼
17	10⅞	AmSeat		.20	1.5	4	2	13½	13½	13½	.....
23⅞	9¾	AndJcb				34	8	11	10¾	10¾	− ¼
14¾	7	Andrea		.60	5.5	14	20	11	10⅜	10⅞	− ⅛
35⅛	12½	AngloE		.27	2.1	4	80	13¼	12¾	12⅞	− ⅜
25	12	ApIDta				12	22	21	20⅜	20⅞	+ ⅜
20½	7⅞	ArgoPtr				22	177	8½	8⅛	8⅛	− ¼
5⅛	2¾	Armtrn				3	18	4⅜	4¼	4⅜	.....
8⅞	5⅜	ArrowA		.20	3.2	89	6	6¼	6¼	6¼	.....
16⅞	8	Asamr	g	.40		7	101	9⅜	9⅛	9⅛	− ⅛
18⅛	10⅜	Astrex				10	6	12⅜	12⅛	12⅛	− ⅜
8⅞	7½	AstrDr	n			4	2	7⅞	7¾	7¾	.....
4¾	1⅞	AtlsCM		.08e	4.0	22	101	2⅛	2	2	.....
8⅞	3⅛	Atlas	wt				25	7	6⅜	6⅜	− ⅛
20⅛	11¾	AtlasV	s	.20	1.5	5	51	13⅜	13	13⅜	+ ⅛
9¼	3¾	Audiotr		.16	2.9	12	49	5½	5⅜	5½	+ ⅛
38⅞	26¾	AutoSw		.80	2.7	13	3	30	29⅞	30	+ ⅜
11⅞	8⅜	AVEMC		.54	5.0	7	35	10¾	10⅝	10¾	+ ⅛
			−	B-B-B	−						
16½	15½	BDM	n	.15e	.9	12	2	16¼	16⅛	16¼	+ ⅛
2⅜	1⅛	BRT					10	1½	1½	1½	.....
7⅜	5⅜	Baker		.16	2.2	11	2	7⅛	7⅛	7⅛	− ⅛
8	5½	BaldwS		.32a	5.7	..	12	5⅝	5½	5⅝	.....
28⅜	21⅛	BanFd		4.70e	21.	..	1	22⅜	22⅜	22⅜	+ ¼
11½	3⅜	Banstr	g				58	6⅛	5⅞	6	.....
3⅜	2⅜	Barco		.12	4.4	14	2	2¾	2⅜	2¾	+ ⅛
9	3½	BarnEn					12	3⅞	3¾	3¾	.....
10¼	3¾	BaryRG				12	49	4¼	4	4	.....
22⅞	7⅞	Baruch		.43t	5.2	7	8	8¼	8	8¼	+ ⅛
15⅛	7¾	Beard	n	.08	1.0	10	21	8⅜	8	8	− ⅛
9⅛	6¼	Behiv	n			5	85	6½	6¼	6¼	− ⅛
4	1¾	Beltran		.10	4.7	..	114	2⅛	1⅞	2⅛	+ ⅛
24⅛	14	BnfStA	s	.40	2.1	8	113	19⅜	19	19⅜	+ ⅜
24	14½	BnfStB	s	.40	2.1	8	21	19⅜	19	19⅜	+ ⅜
34¼	19⅞	BrgBr	s	.48	1.9	12	65	26¼	25½	25⅞	+ ⅜
						1 8 14		27			

			−	D-D-D	−						
4¾	2½	DWG		.34t	14.	3	44	2⅝	2½	2½	.....
14½	7½	DaleE	n	.08e	.9	8	40	9⅝	9¼	9⅜	− ⅛
5⅛	2⅜	Damon				..	9	34	5	4⅝	5 + ½
18⅛	7⅜	Damson		.34t	4.1	12	242	8⅜	8⅛	8⅜	+ ½
8⅜	2½	Damsn	wt				3	2⅞	2⅞	2⅞	+ ⅛
20¼	1¼	DataAc		.15e	5.0	3	94	3½	3	3	.....
44½	17⅛	Datapd		.30	1.6	14	127	18¾	17¾	18¼	+ ½
8	5½	Datrm	n			14	8	5⅝	5½	5½	.....
4	1⅞	DeRose				13	7	2¼	2¼	2¼	.....
2¾	1	Decorat					3	1⅛	1⅛	1⅛	− ⅛
22⅜	13¾	DelLab		.60	4.1	6	3	14½	14¼	14½	+ ⅜
6⅜	2½	DesgnJ		.38t	9.8	4	42	3⅞	3¾	3⅞	.....
35⅞	20⅜	Digicon				12	111	24¾	23⅞	24⅜	+ ½
24⅞	12⅜	Dillard		.40	1.8	5	12	23	22⅜	22⅜	− ⅜
21¼	7⅞	DomeP	s				2227	8¾	8½	8⅜	+ ⅛
31½	16¼	Domtr	g	2			17	16¾	16⅜	16⅜	− ⅛
26¾	14⅜	DorGas		.16	.9	12	176	17⅞	17⅜	17⅜	− ½
10¾	7¼	Dghty		.30b	3.8	7	21	8	7⅞	7⅞	.....
8⅜	2¾	Downey		.28	7.0	..	15	4	4	4	.....
23⅞	8½	Dreco	g				39	8¾	8⅜	8¾	.....
25¼	13½	Driller	n			7	32	14¼	13¾	13¾	.....
11⅞	7¾	DrivHr					1	8¼	8¼	8¼	+ ¼
27⅜	18¼	Ducom	n	.70b	3.5	10	18	20	19¾	20	+ ¼
28⅜	15⅜	Dunes	n			11	38	17	16⅜	17	+ ⅜
16	11¾	Duplx		.68	5.1	5	3	13¼	13¼	13¼	+ ⅛
15⅜	9¾	DurTst		.40a	4.0	9	1	10	10	10	.....
11¾	5⅜	Dynictn		.10e	1.2	6	71	8⅜	8¼	8⅜	.....
19⅛	11¼	Dyneer		s.70	5.2	6	1	13⅜	13⅜	13⅜	+ ⅛
			−	E-E-E	−						
8⅜	4¾	EAC		.36	6.0	5	58	6¼	6	6	.....
19	6⅛	EECO		.28	4.1	15	59	6¾	6¼	6¾	.....
20¾	10⅜	EDO		.40b	2.8	11	35	14⅜	14⅛	14⅜	+ ⅛
5⅞	2	ElAudD				..	3	2⅛	2⅛	2⅛	.....
46	21	ElcAm		1.40	6.4	7	2	22	22	22	− ¼
8¾	3¼	ElecSd				..	23	6	3½	3⅜	3½ + ⅛
13⅛	4⅞	Elsinor				..	10	175	6⅝	6½	6⅝ + ⅛
9¾	2⅜	EmpCar		.50	10.	..	55	5¼	5	5	− ¼
14¾	6¼	EnMgt	n			..	10	74	8⅛	7⅜	7¾ − ⅛
12⅛	3⅜	EngMin		.71t	17.	23	151	4⅛	3⅞	4⅛	+ ⅛
19⅜	9	EngyRs		.24	2.5	10	27	9¾	9¼	9½	+ ½
21½	8⅛	EnrSv	n			..	5	77	10¼	9⅝	9⅝ − ⅜

EXPLANATORY NOTES

(For New York and American Exchange listed issues)

Sales figures are unofficial.

The 52-Week High and Low columns show the highest and the lowest price of the stock in consolidated trading during the preceding 52 weeks plus the current week, but not the current trading day.

u - Indicates a new 52-week high. d - Indicates a new 52-week low.

s - Split or stock dividend of 25 per cent or more in the past 52 weeks. The high-low range is adjusted from the old stock. Dividend begins with the date of split or stock dividend.

n - New issue in the past 52 weeks. The high-low range begins with the start of trading in the new issue and does not cover the entire 52-week period.

g - Dividend or earnings in Canadian money. Stock trades in U.S. dollars. No yield or PE shown unless stated in U.S. money.

Unless otherwise noted, rates of dividends in the foregoing table are annual disbursements based on the last quarterly or semi-annual declaration. Special or extra dividends or payments not designated as regular are identified in the following footnotes.

a - Also extra or extras. b - Annual rate plus stock dividend. c - Liquidating dividend. e - Declared or paid in preceding 12 months. i - Declared or paid after stock dividend or split up. j - Paid this year, dividend omitted, deferred or no action taken at last dividend meeting. k - Declared or paid this year, an accumulative issue with dividends in arrears. r - Declared or paid in preceding 12 months plus stock dividend. t - Paid in stock in preceding 12 months, estimated cash value on ex-dividend or ex-distribution date.

x - Ex-dividend or ex-rights. y - Ex-dividend and sales in full. z - Sales in full.

wd - When distributed. wi - When issued. ww - With warrants. xw - Without warrants.

vj - In bankruptcy or receivership or being reorganized under the Bankruptcy Act, or securities assumed by such companies.

thousands of firms and markets is made fully visible.

The *Journal* is not the only source of business information. Such business magazines as *Business Week, Forbes, Barron's,* and *Fortune* also offer useful information. And detailed financial reports by Moody's and Standard & Poor can be found in most college libraries. But the *Journal* is a remarkable compilation of price and financial information, which gives the reader direct contact with company actions and outcomes.

Table B

Cash Prices

Thursday, February 25, 1981
(Quotations as of 4 p.m. Eastern time)

FOODS	Thurs.	Wed.	Yr. Ago
Flour, hard winter KC cwt	$10.25	$10.35	$10.85
Coffee, Brazilian, NY lb	n1.50	1.50	1.23
Cocoa, Accra NY lb	z	z	z
Potatoes, rnd wht, 50 lb, NY del	y z	z	6.75
Sugar, cane, raw NY lb del	.1731	.1691	.2511
Sugar, cane, ref NY lb fob	.2860	.2910	.3610
Sugar, beet, ref Chgo-W lb fob	.2840	.2840	.3390
Orange Juice, frz con. NY lb	b1.225	1.25	1.32
Butter, AA, Chgo., lb.	1.47	1.47	1.47
Eggs,Lge white, Chgo doz.	.72½-.73	.71½-.72	.66¾
Broilers, Dressed "A" NY lb	x.5060	.5026	.5105
Beef, 700-900 lbs. Midw lb fob	1.04	1.04	.95-.96
Pork Loins, 14 down Mdw lb fob	.96	.98	.91
Hams, 14-17 lbs., Midw lb fob	.84	.81¾	.65½-.66
Pork Bellies, 12-14lbMdw lb fob	.66	.66	.45½
Hogs, Sioux City avg cwt	e48.20	48.80	41.05
Hogs, Omaha avg cwt	e48.10	48.60	41.20
Steers, Omaha choice avg cwt	64.75	64.75	61.75
Steers, Sioux City ch avg cwt	65.75	65.75	61.38
Feeder Cattle, Okl Cty. av cwt	e65.62	65.37	73.00
Pepper, black NY lb	a.78	.78	.79
GRAINS AND FEEDS			
Alfalfa Pellets, dehy. Neb., ton	81.00	81.00	106.-107.
Barley, top-quality Mpls., bu	2.90-3.15	2.90-3.15	3.65-4.0
Bran, (wheat middlings)KC ton	n72.50	74.50	85.50
Brewer's Grains, Mlwke, ton	81.00	82.00	93.00
Corn, No. 2 yellow Chgo., bu bl	2.60¼	2.602	3.47½
Corn Gluten Feed, Chgo., ton	113.67	113.67	114.00
Cottonseed Meal, Memphis, ton	n140.00	143.75	196.20
Flaxseed, Mpls., bu	n7.65	7.65	8.15
Hominy Feed, Ill., ton	67.00	67.00	99.00
Linseed Meal, Mpls., ton	n150.00	151.00	145.00
Meat-Bonemeal 50%-pro, Ill.,ton	205.00	205.00	230.-235.
Oats, No. 2 milling, Mpls., bu	2.12-2.20	2.10-2.20	2.22-2.25
Rice, No. 2 milled fob Ark. cwt	18.0-19.0	18.0-19.0	25.0-27.0
Rye, No. 2 Mpls., bu	3.95	4.00	3.85
Sorghum, (Milo), No. 2 Gulf cwt	5.07	5.18	6.50
Soybean Meal, Decatur, Ill. ton	187.00	187.00	204.00
Soybeans No1 yellow Chgo bu bl	6.06¾	6.04¼	7.25¼
Sunflower Seed, No. 1 Mpls. cwt	n12.00	12.00	11.35
Wheat, Spring 14%-pro Mpls. bu	4.17	4.13	4.69
Wheat, amber durum, Mpls., bu	4.25-4.90	4.25-4.90	6.15-7.75
Wheat No2 soft red Chgo bu bl	3.44¼	3.43¾	4.16¼
Wheat, No. 2 hard KC, bu	4.14¾	4.19¼	4.46¾
FATS AND OILS			
Coconut Oil, crd, N. Orleans lb	z	z	.25¼
Corn Oil, crd wet mill, Chgo. lb	.25¼	c25¼-¾	.23½
Corn Oil, crd dry mill, Chgo. lb	n.26	.26	.24
Cottonseed Oil, crd Miss Vly lb	.19	.19	.24½
Grease, choice white, Chgo lb	.16¼-½	.16¼	.17¾
Lard, Chgo lb	a.22½	.22½	.19½
Linseed Oil, raw Mpls lb	.29	.29	.32
Palm Oil, Neutral, N.Y. lb	z	z	.28¼
Peanut Oil, crd, Southeast lb	n.30	.30	.38
Soybean Oil, crd Decatur, lb	.1778	.1791	.2244
Tallow, bleachable, Chgo lb	.16¾-.17	.16¾-.17	.17¾
Tallow, edible, Chgo lb	.19¼	.19¼	.19½
FIBERS AND TEXTILES			
Burlap, 10 oz. 40-in. NY yd	n.2295	.2285	.2590
Cotton 1 1-16 in str lw-mid Mphs lb	.5977	.5932	.8612
Print Cloth, cotton, 48-in NY yd	s.70	.70	.73
Print Cloth, pol/cot 48-in NY yd	f.48	.48	.57
Satin Acetate, NY yd	.62	.62	.60

a-Asked. b-Bid. c-Corrected. d-Dealer market. e-Estimated. g-f.o.b. harbor barge. Source: Oil Buyers' Guide. i-To arrive by rail within 30 days. k-Dealer selling price in lots of 40,000 pounds or more. f.o.b. buyer's works. n-Nominal. p-Producer price. r-Day's trading range. s-Thread count 78x76. t-Thread count 78x54. x-Less than truckloads. y-Long Island origin; varies seasonally. z-Not quoted.

Table C

	Open	High	Low	Settle	Change	Lifetime High	Lifetime Low	Open Interest
- GRAINS AND OILSEEDS -								
CORN (CBT) - 5,000 bu.; cents per bu.								
Mar	259½	259¾	256½	256¾	– 3½	406½	253	28,358
May	272	272¼	270	270¼	– 2¼	410¾	262½	39,409
July	281	282¼	280	280¾	– 1	399	267¾	30,289
Sept	285¼	286¾	284¼	286	...	388½	268¾	34,370
Dec	292	293¾	290½	292¾	...	345½	271	24,261
Mar83	305¾	307¾	305¼	307	...	320¾	294	4,138
Est vol 62,532; vol Wed 42,686; open int 133,329, – 1,667.								
OATS (CBT) - 5,000 bu.; cents per bu.								
Mar	206½	208	205½	205½	– 1	239	183½	2,490
May	197	199	196½	197	+ ¾	231½	177½	2,432
July	186½	188	185½	186½	+ ¾	207	168½	2,087
Sept	181½	182½	180½	181	– ½	204¾	167	678
Dec	186	187	184¾	186½	+ ¾	199½	182	187
Est vol 1,918; vol Wed 1,528; open int 7,874.								

CATTLE - LIVE (CME) - 40,000 lbs.; cents per lb.								
Apr	65.20	65.30	64.75	65.00	– .55	72.40	53.50	25,156
June	64.15	64.20	63.65	63.80	– .85	72.30	54.75	16,218
Aug	61.80	61.80	61.40	61.65	– .47	66.85	54.30	6,650
Oct	60.00	60.00	59.40	59.60	– .42	65.90	53.70	3,151
Dec	60.02	60.02	59.60	59.92	– .35	64.65	54.90	1,032
Feb83	60.05	60.05	59.80	59.80	– .25	60.90	58.45	38
Apr	60.25	60.50	60.25	60.50	no comp	60.50	60.25	0
Est vol 14,554; vol Wed 16,609; open int 51,174, + 521.								

GOLD (CMX) - 100 troy oz.; $ per troy oz.								
Mar	367.00	367.00	367.00	365.70	– 2.30	410.00	360.50	390
Apr	371.00	373.10	368.20	369.80	– 2.60	898.00	362.70	44,556
June	380.00	381.00	376.50	377.90	– 2.60	925.00	371.00	25,165
Aug	387.80	389.50	384.30	386.20	– 2.60	887.00	379.00	14,549
Oct	395.00	397.00	395.00	394.70	– 2.50	842.00	387.00	16,468
Dec	405.30	406.00	401.50	403.50	– 2.40	666.50	396.00	11,057
Feb83	410.50	410.50	410.50	412.50	– 2.30	642.00	405.00	17,644
Apr	421.60	422.80	421.60	421.60	– 2.20	604.00	415.00	9,310
June	...	...	...	430.90	– 2.10	596.00	425.00	1,649
Aug	...	...	...	440.20	– 2.00	515.50	451.50	619
Oct	...	...	...	449.70	– 1.90	500.00	449.50	466
Dec	...	...	...	459.30	– 1.80	495.00	381.60	38
Est vol 35,000; vol Wed 33,670; open int 141,911, – 1,747.								

HEATING OIL NO. 2 (NYM) - 42,000 gal.; $ per gal.								
Mar	.8650	.8885	.8650	.8850	+ .0276	1.1385	.8135	4,000
Apr	.7890	.8015	.7875	.8000	+ .0180	1.1276	.7760	6,204
May	.7780	.7810	.7735	.7792	+ .0089	1.1300	.7675	3,030
June	.7845	.7860	.7795	.7835	+ .0057	1.0800	.7730	1,964
July	.7900	.7940	.7870	.7915	+ .0064	1.0800	.7825	1,572
Aug	.8025	.8050	.8020	.8045	+ .0060	1.0850	.7975	368
Sept	...	...	...	.8230	+ .0070	1.0850	.8100	81
Oct	.8280	.8415	.8280	.8350	+ .0100	1.0950	.8250	367
Nov	.8560	.8560	.8560	.8500	...	.8560	.8560	0
Dec	.8625	.8680	.8625	.8680	+ .0080	.9475	.8525	143
Est vol 7,000; vol Wed 6,375; open int 17,729, + 174.								

SILVER (CMX) - 5,000 troy oz.; cents per troy oz.								
Mar	799.0	801.5	786.0	788.5	– 9.3	2830.0	779.0	5,307
Apr	810.0	810.0	802.0	799.0	– 11.0	835.0	787.0	21
May	820.0	822.5	804.0	809.0	– 10.0	2895.0	799.0	14,136

CBT - Chicago Board of Trade; CME - Chicago Mercantile Exchange; CMX - Commodity Exchange, New York; CSCE - Coffee, Sugar & Cocoa Exchange, New York; CTN - New York Cotton Exchange; IMM - International Monetary Market at CME, Chicago; KC - Kansas City Board of Trade; MPLS - Minneapolis Grain Exchange; NOCE - New Orleans Commodity Exchange; NYFE - New York Futures Exchange, unit of New York Stock Exchange. NYM - New York Mercantile Exchange; WPG - Winnipeg Commodity Exchange.

Table D

Foreign Exchange

Thursday, February 25, 1982

The New York foreign exchange selling rates below apply to trading among banks in amounts of $1 million and more, as quoted at 3 p.m. Eastern time by Bankers Trust Co. Retail transactions provide fewer units of foreign currency per dollar.

Country	U.S. $ equiv. Thurs.	U.S. $ equiv. Wed.	Currency per U.S. $ Thurs.	Currency per U.S. $ Wed.
Argentina (Peso)				
Financial	.000100	.000100	10025.00	10025.00
Australia (Dollar)	1.0770	1.0786	.9285	.9271
Austria (Schilling)	.0601	.0602	16.64	16.62
Belgium (Franc)				
Commercial rate	.023047	.02308	43.389	43.32
Financial rate	.021282	.02147	46.989	46.57
Brazil (Cruzeiro)	.00733	.00733	136.41	136.41
Britain (Pound)	1.8365	1.8335	.5445	.5454
30-Day Forward	1.8374	1.8345	.5442	.5451
90-Day Forward	1.8415	1.8387	.5430	.5439
180-Day Forward	1.8470	1.8439	.5414	.5423
Canada (Dollar)	.8191	.8199	1.2208	1.2196
30-Day Forward	.8187	.8198	1.2214	1.2198
90-Day Forward	.8181	.8198	1.2223	1.2198
180-Day Forward	.8171	.8190	1.2238	1.2210
China (Yuan)	.5497	.5498	1.8190	1.8190
Colombia (Peso)	.0166	.0166	60.07	60.07
Denmark (Krone)	.1259	.1259	7.94	7.9418
Ecuador (Sucre)	.0404	.0404	24.75	24.75
Finland (Markka)	.0210	.0210	4.5238	4.5243
France (Franc)	.1654	.1658	6.0460	6.0325
30-Day Forward	.1654	.1658	6.0470	6.0325
90-Day Forward	.1651	.1655	6.0550	6.0405
180-Day Forward	.1644	.1646	6.0840	6.0755
Greece (Drachma)	.0164	.0163	61.15	61.39
Hong Kong (Dollar)	.1699	.1696	5.8875	5.8975
India (Rupee)	.1082	.1083	9.24	9.23
Indonesia (Rupiah)	.00154	.00154	648.00	648.00
Ireland (Pound)	1.4925	1.4985	.6700	.6673
Israel (Shekel)	.0569	.0569	17.59	17.59
Italy (Lira)	.000787	.000788	1271.00	1269.00
Japan (Yen)	.004244	.004272	235.65	234.10
30-Day Forward	.004275	.004306	233.90	232.25
90-Day Forward	.004336	.004364	230.65	229.15
180-Day Forward	.004419	.004452	226.30	224.60
Lebanon (Pound)	.2041	.2041	4.9000	4.900
Malaysia (Ringgit)	.4336	.4344	2.3065	2.3020
Mexico (Peso)	.0267	.0256	37.50	39.00
Netherlands (Guilder)	.3840	.3846	2.6040	2.600
New Zealand (Dollar)	.7880	.7880	1.2690	1.2690
Norway (Krone)	.1668	.1666	5.9950	6.00
Pakistan (Rupee)	.0976	.0976	10.245	10.245
Peru (Sol)	.00185	.00185	540.97	540.97
Philippines (Peso)	.1207	.1207	8.285	8.285
Portugal (Escudo)	.01445	.0147	69.19	68.19
Saudi Arabia (Riyal)	.2925	.2925	3.4185	3.4185
Singapore (Dollar)	.4751	.4753	2.1050	2.1040
South Africa (Rand)	1.0285	1.0285	.9723	.9723
South Korea (Won)	.0014	.0014	708.00	708.00
Spain (Peseta)	.00973	.0097	102.78	102.75
Sweden (Krona)	.1732	.1731	5.7722	5.7779
Switzerland (Franc)	.5318	.5329	1.8805	1.8765
30-Day Forward	.5353	.5363	1.8680	1.8645
90-Day Forward	.5416	.5418	1.8465	1.8455
180-Day Forward	.5497	.5500	1.8190	1.8180
Taiwan (Dollar)	.0270	.0270	37.00	37.00
Thailand (Baht)	.0435	.0435	23.00	23.00
Uruguay (New Peso)				
Financial	.0847	.0847	11.80	11.80
Venezuela (Bolivar)	.2329	.2329	4.2937	4.2937
West German (Mark)	.4218	.4224	2.3710	2.3675
30-Day Forward	.4235	.4241	2.3612	2.3577
90-Day Forward	.4271	.4275	2.3413	2.3394
180-Day Forward	.4320	.4325	2.3148	2.3120
SDR	1.12952	1.13300	.885334	.882613

Special Drawing Rights are based on exchange rates for the U.S., West German, British, French and Japanese currencies. Source: International Monetary Fund.

Table E

Money Rates

Thursday, February 25, 1982

The key U.S. and foreign annual interest rates below are a guide to general levels but don't always represent actual transactions.

PRIME RATE: 16½%. The base rate on corporate loans at large U.S. money center commercial banks.

FEDERAL FUNDS: 13¾% high, 13% low, 13¾% near closing bid, 14% offered. Reserves traded among commercial banks for overnight use in amounts of $1 million or more. Source: Mabon, Nugent & Co., N.Y.

DISCOUNT RATE: 12%. The charge on loans to member commercial banks by the New York Federal Reserve Bank.

CALL MONEY: 14% to 15%. The charge on loans to brokers on stock exchange collateral.

COMMERCIAL PAPER: placed directly by General Motors Acceptance Corp.: 13⅜%, 30 to 270 days.

COMMERCIAL PAPER: high-grade unsecured notes sold through dealers by major corporations in multiples of $1,000: 13¾ to 90 days.

CERTIFICATES OF DEPOSIT: 13¼%, one month; 13½%, two months; 13.90%, three months; 14⅛%, six months; 14¼%, one year. Typical rates paid by major banks on new issues of negotiable C.D.'s, usually on amounts of $1 million and more. The minimum unit is $100,000.

BANKERS ACCEPTANCES: 13.40%, 30 days; 13.40%, 60 days; 13.40%, 90 days; 13.35%, 120 days; 13.20%, 150 days; 13.15%, 180 days. Negotiable, bank-backed business credit instruments typically financing an import order.

EURODOLLARS: 14⅜% to 14¼%, one month; 14⅜% to 14½%, two months; 14⅜% to 14¼%, three months; 14 15/16% to 14 13/16%, four months; 15 1/16% to 14 15/16%, five months and six months. The rates paid on U.S. dollar deposits in banks in London, usually on amounts of $100,000 or more. The higher rate for each maturity is LIBOR, the London Interbank Offered Rate.

FOREIGN PRIME RATES: Canada 16½%; Germany 13%; Japan 6.90%; Switzerland 8%; Britain 13.50%. These rate indications aren't directly comparable; lending practices vary widely by location. Source: Morgan Guaranty Trust Co.

TREASURY BILLS: Results of the Monday, February 22, 1981, auction of short-term U.S. government bills, sold at a discount from face value in units of $10,000 to $1 million: 12.430%, 13 weeks; 12.695%, 26 weeks.

SAVINGS RATES: on instruments offered to individuals; minimum amounts vary. Money market fund-a 13.99%; six month money market certificate, 13.958%; 30-month savings institution small-saver certificate (accounts 2½ to less than 4 years) b-14.80%; one-year "all savers" tax-exempt certificates, 10.79%; savings institution passbook deposit-c 5.5%; U.S. savings bond, 9%.

a-Annualized average rate of return after expenses for past 30 days on Merrill Lynch Ready Assets Trust, the largest of such funds; this isn't a forecast of future returns. b-Commercial bank rate. Savings and loan associations and savings banks are permitted to pay .25% more than commercial banks. c-Commercial banks are limited to paying .25% less than savings and loan associations and savings banks.

which the partners are neither horizontal nor vertical. (Chapters 11 and 12 present these mergers' effects on competition.)

Since 1960, most mergers have been conglomerate in nature. This is mainly because antitrust policies have prevented most horizontal and some vertical mergers. Conglomerate mergers appear to be affecting the shape of American business by creating firms that are more diverse than before. Probably most of the products that you buy are made by companies that are really just branches of other larger firms.

Most economists regard conglomeration as interesting but of only secondary importance in analyzing firms' main decisions and outcomes. The economic logic of the firm's investment, production, and pricing are largely the same whether the firm is independent or owned by a larger enterprise.

Other types of firms: Public enterprises, nonprofit firms, and cooperatives

Much of this chapter's content applies not only to the private profit-making firms, but also to public, nonprofit, and cooperative firms.

All firms need to make efficient choices about inputs, outputs, and investments. But these other types of firms differ from private enterprises in that profit is not the single motive for their policies and actions. They usually have social goals as well. Some of these firms seek to supply goods to needy people at low prices. Others provide important services that no private firm could supply at a profit. Still other firms provide "utility" services (such as municipal electric systems and the U.S. Postal Service) for which private operators, having a monopoly position, might charge too high a price.

Taken together, these nonprivate firms are a diverse and important group of enterprises, covering nearly one fourth of all U.S. economic activity. Economists, however, have given them little study, focusing instead on private enterprise.

Public enterprises are found in all sizes at national, regional, state, and local levels. Their conditions are treated in Chapter 13.

Not-for-profit enterprises also include a great variety of firms. They are "owned" by charitable groups and often have some special social purpose. Examples include most hospitals, private schools, and colleges, the Red Cross system, city orchestras and cultural centers, and many day-care centers. Many of these units sell their services, but most rely heavily on contributions. Some struggle along always short of funds; others enjoy ample financing and rapid growth.

Cooperatives are enterprises owned by their customers or suppliers. Millions of farmers sell their crops, livestock, and milk through farm cooperatives, and they buy much of their supplies from them too. In the retail sector, cooperative food stores proliferated in the 1970s. In all cases, the cooperative enterprise tries to cover its costs with sales revenue, and it channels its "profits" back to its owners (customers or suppliers).

There are other types of enterprises, even more uncommon, such as worker-owned firms. But virtually all business in the United States is conducted by private firms, public firms, nonprofit enterprises, and cooperatives.

The enterprise

The essence of the enterprise (or firm) is clear and simple: "Enterprise" is just a term for any unit where people produce a good or service. The enterprise may consist of one local plant (factory or office) or more, on up to many hundreds of plants. The corner drugstore is an enterprise, and so are General Motors Corporation, the

Chicago White Sox, the hospital where you were born, and your own college.

Choices and outcomes

Whatever its size or form, the enterprise has the basic task of combining inputs in a production process to create output. These choices *compare benefits with costs,* as do other economic choices. *Inputs* are costs, while *outputs* provide benefits in the form of the sales revenue. The manager carries each choice to the point where its marginal benefits just equal its costs. And in total, the manager tries to maximize the excess of benefits (revenues) over costs.

The basic process is shown in Figure 3. The firm makes choices: what good or service to produce, what kinds and amounts of inputs to use, what quantities of output to make, and what price to set for its output. Not only are these choices numerous, but often the range of choice is also very wide. Once the firm decides and proceeds, the resulting outcomes are quite definite.

But behind each definite outcome lies *a range of alternative values that could have been selected.* It is the economist's task to explain how the firm chooses among all of these alternatives.

The firm plays two main roles in its task of turning inputs into outputs. **First, it is an *owner and manager of real capital,* which is used in production.** In this role, the firm raises funds and then invests them in building up its physical plant and equipment. This activity is reflected in the firm's yearly balance sheet, which shows the *stock* of value contained in the capital.

Second, *the firm is a producer, using labor and capital to process materials into outputs.* The *flow* of costs of inputs and the revenues from outputs is presented in the firm's yearly income statement.

The cost paid out for each input is simply its price times the quantity used in a given period. ***Total cost*** paid out to inputs by the firm is then simply the sum of all these separate input costs: the labor, raw materials, capital, and so on. On the revenue side, ***total revenue*** is equal to the price of

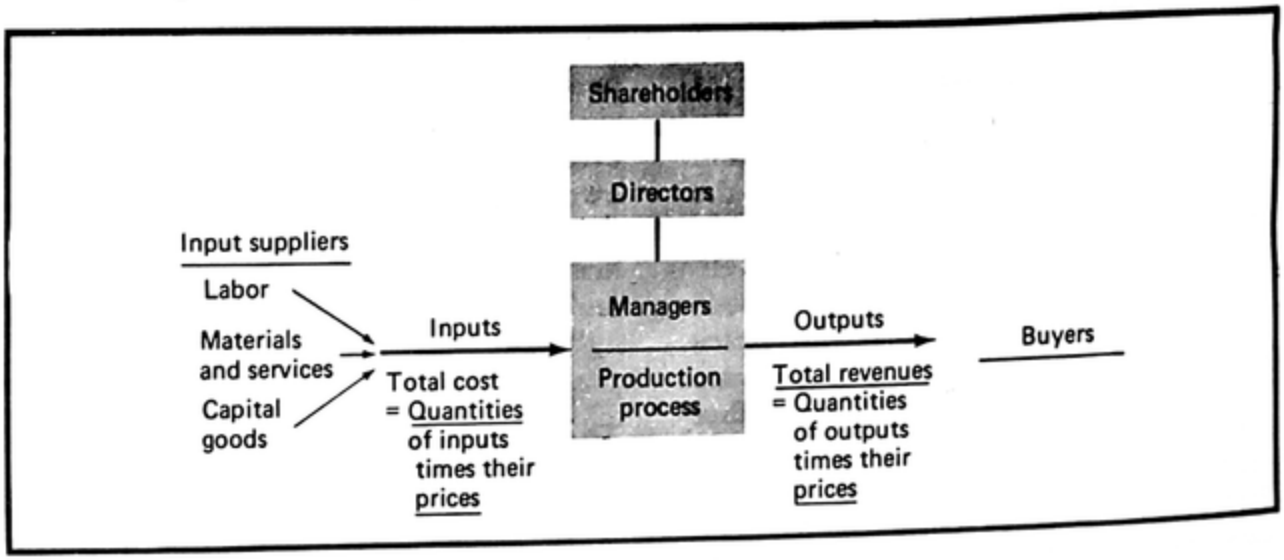


Figure 3 The basic economic elements of an enterprise

The firm uses inputs to produce outputs via some process of production. Each input has a price and is used in a definite quantity during the period. The price times the quantity is the cost of that input. The input costs together make up total cost. Similarly, each output has a price and is sold in a definite quantity during the period. Each output's sales revenue is its price times its quantity. Together, the outputs' revenues make up total revenue.

each unit of output times the quantity of output sold. Total revenue is often called sales revenue or gross receipts.

Net income or profit is simply the revenue left over after all required costs are accounted for:

$$\text{Net income} = \text{Total revenue} - \text{Total cost}$$

$$\text{Net income} = \begin{pmatrix}\text{The sum of all the}\\ \text{output values}\end{pmatrix} - \begin{pmatrix}\text{The sum of all}\\ \text{the input costs}\end{pmatrix}.$$

If the firm is privately owned, its managers' main aim is to maximize the firm's profits. They try to use the "right" mix of inputs in an efficient production process to produce the most profitable amount of output. The firm's manager may have a degree of choice in choosing the price-quantity combination. Or, instead, the output prices may be set by market competition, so that the manager is only free to choose the right amount of output, given the price. Although the managers try to keep revenue up and costs down, profits are not assured. Successful performance depends crucially on skillful estimation of consumer preferences and cost conditions. Wrong choices about what to produce and how to produce it will be penalized by losses. Input and output choices, then, are motivated by the search for profits.

Production is therefore merely an outcome of the pursuit of profits. If the markets are working well, the pursuit of profits will lead private firms to provide the array of products that consumers want, at minimum costs.

Inputs, outputs, and production

Inputs As their name indicates, inputs include all items put into the production process. The three traditional main classes of inputs are labor, capital, and land (which includes all natural resources).

Labor is the application of human effort, both in physical force and mental skills. The effort is provided by workers, selling their services by the hour, day, week, or month. Labor comes in many types, ranging from simple actions like digging or fastening, on up to complex professional skills. Managers are also a form of labor input. These various kinds of labor are priced at varying values. Chapter 15 analyzes labor and wages in detail.

Capital is the stock of productive assets created by past investments. It includes buildings, equipment, roads, and any other improvement to natural conditions such as dredging a harbor or clearing stones from a field. Capital increases productivity by enlarging what labor can do. Thus, a hammer and a 20-ton drop forge both increase the ability of workers to bend metal. Computers increase calculating abilities, and jet airplanes increase the speed of travel. Capital is discussed more fully in Chapter 16.

Land is a broad term for all natural resources and raw materials. It includes oil, minerals, forests, farmland, and even fishing shoals. Resources range from nonrenewable ones (ores, coal) to virtually inexhaustible ones (air, solar energy).

Besides these three classes of inputs, there are goods that firms sell to other firms. Thus, a steel plate is an output of a steel company but an input to a machinery firm. Such ***intermediate goods*** fall into three classes of their own. One is finished goods sold to firms for use in *their* products—for example, tires and batteries sold to automobile companies. Second is semifinished goods, such as iron slabs, industrial chemicals, paper, rubber, or flour, which will undergo further processing. Third is services, as distinct from physical products. Examples are electricity, insurance, advertising, and the transport of goods.

Such varied inputs provide large degrees of choice to firms. They must decide not only which inputs to use but also how much of each one to buy. The decisions depend on two fundamental sets of conditions—*technology* and *the prices of the inputs*. Technology, the "state of the art," governs how inputs can best be combined. For example, aluminum, steel, and plastic can be used in various parts of automobiles. Each has certain technical advantages of strength, lightness, and flexibility.

But technology alone does not decide the choices among these three. The other basic determinant is the *prices of the inputs*. Thus, if aluminum's price rises sharply compared to plastic's price, then plastic may extensively replace aluminum to minimize the total cost of an output.

Outputs The two main categories of outputs are goods and services. *Goods* are physical things, such as a gallon of gas, a ton of bricks, or a box of corn flakes, and can be stored. *Services* are less tangible and often have no lasting physical form—TV repair, insurance, legal advice, and a blood test, for example. Often goods and services are mingled: A new car may have a guarantee of certain repair services.

Outputs are usually defined by their *location* or *timing* and by their physical features. As for location, a ton of coal delivered to your door differs, economically, from a ton located at the mouth of the mine. Strawberries available at the grocery store differ from those still unpicked at the farm. As for timing, a plumber's repair call made on Sunday night differs from one made on Monday morning. In all these cases, both (1) the cost and (2) the nature of the good itself can be different, despite the seeming uniformity.

Production is any process that converts inputs into outputs, as shown in Figure 3. Production occurs in all manner of plants, under a great variety of conditions. There are dark clanging mills, spotless electronic assembly lines, deep mines, bustling stores, and countless other productive scenes.

The diversity of production techniques is quite clearly reflected in the size of firms. In some industries, such as steel, automobiles, and the production of electricity, production techniques require large size for efficiency. For example, the smallest auto manufacturer, AMC, hires 27,000 workers and owns $1.1 billion worth of capital. In the retailing industry, in contrast, corner grocery stores coexist with nationwide sellers such as A & P.

Simple accounting

Each year the diversity of firms' activities is distilled into standard accounting measures. As we have already noted, there are two parts to such measures: the income statement and the balance sheet. These give the precise accounting data about what the firm has done. The ***income statement*** sums up the results of the firm's production choices, while the ***balance sheet*** covers the firm's management of its physical and financial assets. You can gain practice in interpreting such accounts by looking up real companies' accounts in *Moody's Industrial Manuals*, comparing their entries with those discussed here.

Note that we are discussing *accounting* costs and profits. The more rigorous *economic* concepts of costs and profits will be treated in Chapters 8 and 9. The two versions are related, but the economic analysis is deeper and therefore different in certain parts, especially in the definitions of costs and profits.

Income statements The top line in an income statement represents the firm's sales revenue, as shown in the sample statement in Table 2. The next lines cover the various

Table 2 ***Yearly accounts for a typical firm (millions of dollars)***

Income Statement

	1982	1981		1982	1981
Sales Revenue	*219*	*191*	*Earnings before Tax*	*41*	*27*
Cost and Expenses	*178*	*164*	*Taxes on Earnings*	*19*	*13*
Labor and materials	85	80	*Net Income after Taxes*	*22*	*14*
Materials	53	48	*Dividends*	*8*	*6*
Services	19	16	*Retained Earnings*	*14*	*8*
Depreciation	8	7			
Interest expense	13	13			

Balance Sheet

Assets	1982	1981	Claims on Assets	1982	1981
Current Assets	*51*	*47*	*Current Liabilities*	*27*	*25*
Financial	43	40	Accounts and rates payable	20	19
Inventories	8	7	Other	7	6
Gross Plant and Equipment	*290*	*268*	*Long-term Liabilities: Debt*	*130*	*128*
Less depreciation	62	54	*Stockholders' Equity*	*122*	*108*
Net Plant and Equipment	*228*	*214*	Common stock (22 million shares at $1 par)	22	22
			Retained earnings	100	86
Total Assets	*279*	*261*	*Total Claims on Assets*	*279*	*261*

costs, which the firm must pay from its revenues. Most accounts lump the operating or production costs together; they include wages and salaries, materials and services. The wages and salaries paid to labor are usually the largest single cost, averaging about 50 percent of all costs. Materials are usually next in size. After operating costs come two *costs of capital.* The first is *depreciation,* representing the yearly wearing out and obsolescence of machinery and buildings. This decline in the value of the capital is made good by setting aside funds for the replacement of the capital. The second cost of capital is *interest* on the company's debt (its bonds and borrowings).

The difference between revenues and costs is earnings (or accounting profit) before tax. From those earnings, the federal tax on profits usually takes about 45 percent. The ***after-tax profit*** is the company's yearly financial payoff for its ownership and production actions. The profit can be large, small, or negative. Since profit usually differs from year to year, it is necessary to take the *average* profit over a period of time to determine the firm's true profitability.

The firm usually pays out some of the accounting profits to shareholders as ***dividends.*** The paying out of dividends is optional; the firm can omit them, change them, keep them steady, do whatever it thinks best from year to year. The remainder of the profits, about two thirds on average, is then kept by the firm as ***retained earnings.*** These funds can be used for expansion or other actions that will increase the value of the firm.

Balance sheets The firm has productive assets, which appear on the left-hand side of its balance sheet. The firm has issued paper securities in the form of stocks and bonds to the people who gave it the money to buy these assets. These paper assets are liabilities to the firm, since they represent claims

against its wealth. They appear on the right-hand side of the firm's balance sheet. The stocks are the owner's claim on the firm's assets. By accounting methods, the total values in the asset and claims sides of the balance sheet are always equal.

Assets include two categories: current assets and long-term assets. They are listed in decreasing order of "liquidity," which is the length of time ordinarily needed to convert them into cash. *Current assets* are mainly cash, accounts receivable, and inventories. Only cash represents actual money. Receivables are the amounts owed to the firm. Inventories are the raw materials, work in process, and finished goods available for sale. The receivables and inventories will presumably be converted into cash eventually.

Fixed assets are the real plant and equipment that the firm has built up over the years. They include machinery, buildings, land, and any other valuable and lasting capital that is used in production. They are listed first at their *gross original value,* which is the sum of all the prices paid for the items when they were acquired. There is also the sum of *depreciation* accrued in order to offset the deterioration of the capital. The difference is the net accounting value of the firm's fixed capital, called *net plant and equipment.*

The ***claims against these assets*** are of two types. First, ***liabilities*** are amounts of money owed by the firm to its bondholders and to others who have lent money to the firm in loans of varying lengths. Liabilities remain constant except as they are directly paid off or added to by more borrowing. Liabilities impose the cost of interest payments, which must be made if the firm is to remain in business. Failure to meet those payments results in insolvency, which, if continued, may lead to bankruptcy.

Second is ***equity*** (or net worth) of the firm: assets minus liabilities equals stockholders' or investors' equity. If the asset values were to decline, that would cause equity to decline, for liabilities remain constant unless changed directly. Therefore, the risk that asset values will decline is borne by the shareholders. If asset values rise, on the other hand, the benefits go to the shareholders.

The accounting value of stockholders' equity arises from two main sources. One is the original money acquired by selling stocks when the firm was created. *Retained earnings* make up the rest of equity. They are simply the sum of all income retained earnings over the years of the firm's existence.

Accounting values for equity represent the owners' stake in the business, but only in accounting terms. ***The actual market value of the firm as judged by investors is determined by the daily buying and selling of the firm's stock in the stock market.*** The stock's price may fluctuate widely. Often the firm's market value moves broadly in line with the book value of its assets and stockholder equity, but there is no direct tie. Indeed, the challenge for management is to deploy the firm's assets so that their value in use—in generating excess profits—will be much greater than their cost. The extra value can be created by good management, luck, monopoly power, innovation, or simply by inflation.

The accounting values for equity do not show these opportunities. Rather, they merely record the sum of past amounts. Typically the retained earnings will be the largest part of total equity. For example, Procter and Gamble's equity was recently $4.2 billion, of which $3.6 billion was retained earnings from earlier decades.

The firm's aim is to have large and growing profits as a return on stockholders' equity, so that it can both pay dividends to the stockholders and build up the business through investment. The stockholders can benefit either way. The divi-

dends give them an immediate reward. The plowing back of retained earnings will increase the firm's capacity and prospects for future profits. That will, in turn, increase the value of the business and cause the firm's stock to be bid up in the stock market. Therefore, retained earnings can give the owners a capital gain in their stock prices, as opposed to the immediate gain they receive from the direct payment of dividends.

Success indicators: Profitability and stock prices

Profitability is the main index of a private firm's economic performance. The company will naturally publicize its other socially attractive activities, such as the number of jobs it creates, its production of high-quality outputs, its exports, innovations, and so on. But these are all secondary to the firm's main goal: to earn a large and increasing flow of profits for its investors.

Profitability is a matter of degree, not of absolute amounts. The simple total of dollar profits is not enough to show how profitable a firm is. A local lumber company with $1 million in profits in a year may have a higher degree of profitability than the largest oil firm, Exxon, with its yearly total profits of over $3 billion. The reason is that *profit as a percentage of capital* or ***rate of return on equity*** is the correct measure of profitability, for that shows how well the firm is managing its owner's capital.

Note that profits as a percentage of capital is *not* the same as profits as a percentage of sales or costs. A bookstore's profits on the textbooks it sells, for example, may be only 3 percent of its sales of those books—"a few pennies on the dollar." Suppose, however, that the bookstore has yearly sales that are ten times as large as its capital (its capital is mainly just the building, shelves, and inventory). Then its 3 percent profit margin on *sales* would be a 30 percent return on its *capital.* That would be a high rate of return on the investment, not a low one. So, once again, ***always appraise profits as a return on capital.***

The simple formula for profitability is:

$$\text{The rate of return} = \frac{\text{Net income after taxes}}{\text{Capital}}$$

$$= \frac{\text{Total revenue} - \text{Total cost and Taxes}}{\text{Invested capital}}.$$

For total invested capital, the usual accounting figure is stockholders' equity. You can easily calculate it for the sample firm's 1982 results in Table 2:

$$\text{The rate of return} = \frac{\$22 \text{ million}}{\$122 \text{ million}}$$

$$= \frac{\$219 \text{ million} - \$178 \text{ million} - \$19 \text{ million}}{\$122 \text{ million}}$$

$$= 18.0 \text{ percent}.$$

This figure is for one year. To judge the firm carefully, you must consider the average profit over some three to five years, so as to even out any odd yearly fluctuations.

Each owner and manager seeks profit rates much higher than the 8 to 10 percent that is the average rate of return. Their nightmare is to run losses. Only by managing production well and keeping costs low and revenues high can the firm's officials produce good profits for the owners.

Stock prices are the other main success indicator for the private firm. Each share of stock offers its owners a chance to get future dividends and capital gains (that is, a rise in the price of the share itself). The firm's managers want to satisfy the investor-owners by making the company prosper, so that (1) dividends will grow and/or (2) the stock price will rise and provide capital gains. The share's price depends on demand and supply in

the stock market. And both supply and demand, in turn, depend on what investors think of the company's performance.

Since most large-scale investors are pretty well informed, they act quickly. Therefore, stock prices usually move swiftly and sensitively. If prospects for the company turn better, then more investors will want to buy shares in the company immediately to be able to share in future benefits. Since fewer investors will want to sell, the increase in investor demand will cause the stock price to rise without delay. Conversely, a downturn in a company's future prospects will cause investors to sell the stock now, before the price goes down. Yet that will quickly cause the stock's price to fall, for sellers will have increased while buyers will have decreased. The only way to sell the stock is to accept a lower price for it. In either case, investors hoping to act before a price change will, by their very actions, bring about that change immediately.

Accordingly, the market value of a stock largely depends both on the firm's *future* prospects and on its current performance in maximizing profits. There are some other influences also, such as the general level of interest rates and broad shifts in average prices for the entire stock market. ***Yet current stock prices are usually a sensitive, quickly adjusting index of investors' judgments about each firm's whole performance, both present and expected.*** Stock prices reflect expected future gains. We will show in Chapter 16 how this discounting feature of the stock market tends to apply steady pressure on firms to maintain their efficiency.

Profits are also a signal for investment. When an industry has high profitability, it is a signal that that industry needs more investment. A high rate of profit shows that the value of the firms' output is well above its cost. Therefore, extra output would be worthwhile because people will pay more than the present level of what it costs to make it. To increase output, one must expand capacity. To expand capacity, one must invest more. High profits are like a green traffic light, signaling more capital to come ahead.

In contrast, financial losses are like a traffic light flashing red, showing a need to reduce investment and to contract capacity. Consumers in the market will not pay enough in sales revenues to cover the cost of the output. Therefore, the output's value is less than its cost. Lower levels of output should be produced, and capacity should shrink. That requires cutting back on investment or even admitting that some of the existing capital has lost its value (in accounting terms, one "writes off" the now-valueless assets). The process of shrinking the amount of capital is logical, since private investors will naturally shun a company that is losing money.

A case study: Starting a new enterprise

To draw together the concepts in this chapter, we will now trace the typical steps involved in starting a new enterprise. We will cast you as the firm's founder.

Choosing what to produce

First the creator of a new firm must perceive an unmet need. The need might be for a new local newspaper, a special hand tool, an electric car, or a housecleaning service. It is wise to avoid crowded markets, where supply is already ample and no firm can currently make high profits. One looks for "new" markets, where the sellers are few and there are good chances for unusually high profit rates. One needs a clear concept of the new good to be offered, a good sense of market realities, skill in organizing, and considerable stamina. You

will be competing both against the established firms and with other ambitious people who decide to enter the market after you.

You will need to plan ahead about the product, its inputs, the method of producing it, where to make it, and the customers for it. The plans for costs, prices, and financing must cover at least the first several years of operations, not just the first few months. Estimating conservatively, the investment must offer at least a 15 percent return on capital, plus a premium (5 or 10 percent more) for the extra risk in the business. Otherwise, no banker or investor group will provide finance—a clear sign that the venture will probably fail.

You narrow the choices down to three possibilities: A restaurant featuring wholesome food, set in an uncrowded location, seems to offer a 40 percent return on capital. A fast-food franchise is available, which might yield a 25 percent return. Or you could organize a "House Care" firm, which performs painting, fixing, cleaning, and so on. It would pay a 30 percent return on investment, but at a higher risk. Assessed objectively, the fast-food franchise is too easy to imitate, and house care would be too hard to supervise (absenteeism, arguments over quality of work, etc.). The restaurant venture offers more security and growth, and so you choose it.

Starting

The first requirement is to obtain capital; the second, to organize production. Recall that these are the two basic functions of a firm: to manage assets and to produce. Your request for financing is turned down by numerous bankers and investors, who consider your undertaking too risky. Finally you raise $500,000 from various investors, including relatives. You incorporate the company and issue shares in the business to the investors. The shares are equity capital; the firm will pay dividends on them when it seems best to do so.

You lease the location, rent the restaurant equipment, and remodel the interior. You hire the staff, design the menu, order supplies, and begin advertising. (The determinants of these input decisions—about equipment, workers, food, and other supplies—will be explained in the next chapter.) Your grand opening occurs one year after your original decision to start a new firm.

The pace of activity is slow at first, because consumers take time to adjust to new products at new prices. The restaurant operates below half of its capacity even at peak mealtimes. Because sales revenues are lower than costs, the firm is losing money. Though initial losses were expected, they are now large enough to strain the firm's finances. You increase advertising, improve the services, and offer price discounts. The rate of operations rises, and soon the restaurant is regularly filled. After one year, the venture is still running losses, as Table 3 shows. There are two major problems: (1) daytime costs are high but traffic is low, and (2) overhead costs are high compared to your small space. There is a need to develop daytime sales and to add space. Your choices are now very risky.

At precisely this moment, a local investor offers $1 million if you will issue shares giving him 75 percent of the firm's stock and thus control of the enterprise. The additional funds would cover the expansion, but it would end your control. Instead, you persuade the original backers to provide the expansion funds. The budget is tight but the restaurant is getting established. It earns $150,000 profit during 1982, which after taxes (including an offset of $50,000 for the first-year loss) is a 10 percent return on the $1 million investment.

Table 3 ***Yearly results for the new restaurant enterprise***

	1981	1982	1983	1984	1985	1986	1987	1988	1989
Workers	15	28	28	28	28	80	80	80	80
Sales volume ($1000)	300	800	900	850	1050	2500	2800	3200	3200
Costs ($1000)	350	650	630	630	700	1900	2100	2100	2100
Accounting profit ($1000)	−50	150	270	220	350	600	700	1100	1100
New investment ($1000)	500	500	0	0	0	2000	0	0	0
Assets ($1000)	500	1000	1000	1000	1000	3000	3000	3000	3000
Profit (before tax) as a % of investment	−10%	15%	27%	22%	35%	20%	23%	37%	37%
Profit (after tax) as a % of investment	−5%	10%	14%	11%	18%	10%	12%	19%	19%

The restaurant is also profitable during 1983, but in 1984 trouble arises. The staff goes on strike for two weeks, and profits decrease. Officials of a restaurant chain offer to buy the firm for $1,050,000 (that is, at a net 5 percent profit to the investors). You persuade the firm's board to reject the offer. In 1985, the finances recover, and you start two more restaurants in other towns at an additional investment of $2 million. They are completed in 1986, and during 1987–1989, the whole firm (with its three plants) is profitable.

The business is now established, despite the crises of the first years. You worked hard and took great risks, for a modest salary. Congratulations! You have succeeded by choosing an excellent location and an attractive style, by applying business sense and skills, and by enjoying the support of your investors. But in 1989, new competition enters, as similar restaurants are opened nearby. You now assess the prospects. Sales are likely to stabilize at about $3.2 million per year. The yearly after-tax profits are $550,000, which gives a 19 percent return on investment. That flow will probably continue into the future.

The firm is now extremely valuable. The $550,000 yearly profit stream is worth perhaps $5 million in the financial market, because buyers would probably pay a capital value of about ten times the level of new income. Indeed, a national restaurant chain does offer to buy the firm for $6 million. The shareholders would all gain, since they had only put in a total of $3 million. By selling your one third of the firm's shares, you would gain $2 million in capital value. The years of hard work would have made you a millionaire. But by selling, you would become just a branch manager, and you might then be fired or demoted, rather than kept on or eventually promoted to an executive position in the national chain. Of course, you could now found another firm, drawing on your experience and financial connections.

This choice is a fork in the road. It arises for many successful small businesses. Which direction would you take?

Lessons

The case study has illustrated six lessons:

1. A firm's output must have a value to buyers that exceeds the costs of producing it, by an amount large enough to pay for the owners' risks and the managers' efforts. When the production costs continue to exceed the revenue, the firm will, and should, stop producing. Once again, the economic

comparison is between benefit and cost.

2. The new firm usually has a difficult start-up period. Only when it is established may profits be expected to flow in.
3. Competing in the market is risky, especially for new firms. The prospect of extra profits can induce people to take those risks. If the risks are greater than the returns, then the investments should *not* be made.
4. The pursuit of commercial success is responsible for much of the unremitting activity of capitalism. Production, jobs, and innovation are side effects of the drive for profits.
5. Failure is easy, especially when planning, financing, and day-to-day operations are not done with utmost care. To start an "independent" business in the "free enterprise" system is usually to undergo strong financial and market pressures. The range of choice is often small, especially at first.
6. The financial market continually assesses a firm's performance and prospects. Success often brings efforts to take over the firm, with or without the consent of its managers. The market for corporate control, for buying and selling whole companies, operates parallel to all of the markets for goods and services.

We will return to this case study occasionally in the next several chapters.

Summary

1. Small and medium-sized enterprises perform most U.S. economic activity. Large firms are important but are concentrated in a few sectors.
2. The firm is the basic unit of production and supply. Each firm has two main functions: It owns and manages assets, and it produces outputs from inputs. Inputs involve various kinds of costs. The outputs are sold at market prices, bringing in sales revenue. Profit (if any) is the excess of revenues over costs.
3. The rate of profit on invested capital shows how profitable the firm is.
4. Sales revenues and all costs are shown in the firm's income statement. Costs include both depreciation of the firm's capital and interest payments. Profits may be paid out to shareholders in dividends or reinvested in the firm.
5. The balance sheet shows the firm's assets and claims on those assets; their totals always equal each other. Assets are current and fixed. Claims on assets include debt and stockholders' equity.

Key concepts

Private Enterprise
Inputs
Outputs
Total cost
Total revenue
Net income (profit)
Production processes
Income statement
Balance sheet
Dividends
Retained earnings
Assets
Claims against assets
Liabilities

Equity
Profitability: rate of return on equity

Questions for review

1. Why do so many new businesses fail?
2. Why has ownership increasingly become divorced from the control of U.S. corporations? What are some possible effects of this situation?
3. Explain why a firm's accounting values for equity can be different from its actual market value.
4. Firm A earned $3 million in profit last year; firm B earned $1.5 million in profit. Is firm A therefore more profitable than firm B?

8

Supply: The Nature of Costs

As you read and study this chapter, you will learn:

- basic concepts of technology, opportunity cost and profit
- the analysis of productivity and costs in the short run, including the law of diminishing returns
- the analysis of long-run productivity and costs

In 1980, there was a spectacular contrast between AT&T, the telephone company, and Chrysler Corporation, the automobile producer. AT&T earned $6,079,000,000 in profits after taxes, while Chrysler lost $1,709,700,000. AT&T was resoundingly prosperous; Chrysler was on the verge of corporate death.

To recover, Chrysler mounted a promotional blitz to increase demand for its cars. But its most direct and desperate actions were internal, as it wielded the knife to cut its costs. Plants were closed, staff was pruned, and production workers were laid off. The surgery was painful but necessary.

Indeed, keeping costs down is every competitive firm's main economic task. Managers will drive their engineering staff to the breaking point to save half a cent on the cost of a mass-produced item, since half a cent may be the difference between a profit and a loss. No firm that wants to succeed in any industry can afford for long to ignore opportunities to reduce its costs. Chrysler's agonies are merely an extreme case of this universal problem.

This chapter is devoted to the study of costs from the economist's point of view. Its main function is to provide the underpinnings of the theory of supply. Its three main sections are devoted to basic concepts, cost variations in the short run, and the determination of cost in the long run.

Basic concepts: Technology, opportunity cost, and economic profit

Technology

Technology is the starting point, the bedrock of cost. It is the *state of the art*, the knowledge about the best techniques of production. In each industry, the current technology defines the firm's alternative choices in using inputs to produce outputs. By choosing the best combinations of inputs, the firm can minimize its costs. Technology exists in every age, but it also evolves. For example, the technology for modern computers and jet aircraft did not exist in 1930, but it did in 1960. Indeed, modern industry differs radically from 19th-century industry, because technology has advanced so far in so many industries.

Technology usually offers a wide variety of choices. Whether you bake bread, smelt iron, or weave cloth, there are many methods of production to choose among. To take a simple case, suppose that a firm wants to machine 5,000 metal parts. It may be able to do this with 10 people and 5 large machines, or 15 people and 12 small machines, or 40 people and 40 hand tools. Which method should it choose?

One important step toward maximizing profits is to make sure that a given level of output is being produced at the lowest possible cost—that is, with the best possible mix of inputs. A firm can hardly be maximizing its profit by producing 2,000 units at $5 per unit if those 2,000 units could be produced at $3 per unit. Thus, any profit-maximizing firm must choose the technology that is the *least-cost* or *economically efficient* method of production for its level of output.

How does a firm find this least-cost or most efficient technology? It must consider both the physical technology of combining inputs (reflecting engineering relations and physical processes) and the relative prices of inputs (which reflect their relative scarcities). At each point in time, with existing technology and input prices, there is a method of production that minimizes costs for a given level of output.

Reasons for variety While the criterion for choosing the most economically efficient technology is straightforward, the process of choice is not. Even firms that produce the same output may choose different technologies or mixes of inputs. For example, a firm's location may influence its costs and therefore input choices. One firm may be located in an area of cheap labor, so it chooses a labor-intensive method of production.* Another firm may be near a low-cost ore deposit, so it chooses a technology that makes much use of ore. A third may have poor access to transportation, so it uses inputs that are close to the plant.

Technological choices also differ because human judgments about the least-cost method of production may differ. After all, technology involves complex choices with uncertain outcomes. One manager may expect labor costs to rise rapidly compared to the price of machinery, so he may choose a technology involving relatively more capital and less labor. By contrast, because another manager

*When a technology uses an input in unusually high proportions, it is expressed as "intensive." Thus, a labor-intensive method uses a high proportion of labor compared to other inputs—for example, the hand picking of farm crops. A capital-intensive method would use large machines operated by only a few people.

may expect capital costs to rise more rapidly, she might choose a technology involving relatively more labor and less capital. To complicate matters further, the known technology or state of the art is constantly changing. The best choice this year may be outdated next year or the year after.

If all of these complications were discussed now, you would be swamped with detailed ifs, buts, and howevers. Instead, it makes sense to begin the analysis simply, concentrating on the most basic principles of cost and supply. Technology is assumed to be fixed, so that definite choices can be made. This is called *static* analysis, and it allows a clearer look at input choices, productivity, and costs at a given time. The best technology is also assumed to be well known, and not uncertain. With these conditions, economic tools can be applied more clearly to show how a firm chooses its inputs. Of course, once you have grasped the basic principles, you can apply them to increasingly complicated situations. For example, a later chapter deals with dynamic issues, involving changes in technology and innovation.

Opportunity cost

You already know that one goal of a firm is to produce its level of output at the lowest cost possible. But what do economists mean by *cost?* Your immediate response might be the dollars that the firm pays out or owes to others. That is certainly important, but it is only part of cost. To see why, start with a very general definition of cost as *the value of inputs used in the process of production.* If you think back to Chapter 2, you will remember that value comes from scarcity. An input has value, then, because it is scarce—if you use the input to produce one good, it is not available to produce something else. An economist, therefore, measures cost by using the concept of *foregone alternatives.* The cost of taking one action is measured in terms of what was given up—in terms of the most highly valued alternative that is sacrificed. And the cost of producing one good measured in terms of foregone alternatives is what could have been produced instead. **When cost is calculated by using this idea of foregone alternatives, it is called *economic cost* or *opportunity cost.***

The concept of economic cost applies to far more than just production. Every action that you take involves an opportunity cost. For example, what is the opportunity cost to you of attending an economics class? Well, first consider all of the alternatives. What could you have done instead? You might have slept an extra hour, studied for another course, listened to a record, jogged, or worked at a part-time job. From your list, pick the most valuable alternative to attending the economics class. That is the opportunity cost of going to class. For you, the opportunity cost might be that foregone hour of sleep. For a classmate, it might be an extra hour studying history. What are you doing Saturday evening? Going to a party or a movie? Sleeping? Studying? Whatever you decide, try to calculate the opportunity cost involved.

Opportunity cost is forward looking, based on a range of choices. The firm compares the alternatives before choosing one of them. The cost of the best alternative is the value that the next best alternative would give. Once the choice is made, the economic content is over, the cost is fixed, and events can then run their course. At the end of the year, actual costs are recorded and rendered in the accounts. In the restaurant example in the last chapter, the decision to expand during 1982 involves complex comparisons of values for the $500,000 of added investment. By 1982, that investment was made: It had become a fixed cost, *not* an opportunity cost.

The *accounting costs are backward looking and rigidly specific.* They are merely the surface numerical outcome of the economic choices that were made earlier. One must keep opportunity cost and accounting cost in separate mental boxes. The meaning of opportunity cost is like a secret password of microeconomics. Economists know it precisely; most other people don't. Rational firms use its content, but they often don't state or apply the concept explicitly.

The economic cost or opportunity cost of production must include a value for all scarce inputs used. That would include all inputs that have an alternative use. Many inputs are bought or hired by a firm. If the market is working properly, the price paid for the input will reflect its value. Part of the cost of the firm can be calculated by simply adding up the dollars paid out by the firm. These are the ***direct costs,*** or ***explicit costs,*** or ***accounting costs*** of the firm. Direct costs include the purchase of raw materials, equipment, wages paid to hired employees, rent, interest, and utilities.

But think of all the scarce inputs that a firm uses, for which there are no market transactions. For example, suppose that you own your own store and put a lot of your own time and money into the business. Certainly your time and money are scarce inputs, since they have many alternative uses. If a value for them is not included in your firm's costs, then the true cost of production is not being given. But how do you estimate a value for your time and money? And what about the cost to your firm of other inputs for which there is no market transaction, such as the use of machinery, depreciation, or the use of a patent, formula, or brand name that you could have sold instead of using yourself?

Remember that value comes from foregone alternatives. Why not use this concept to estimate a value for the scarce inputs for which there is no market transaction? Simply *impute or estimate* a value for these inputs equal to the return they would get in their best or highest-paying alternative use. This measure can be called ***imputed cost*** or ***implicit cost.*** For example, you might believe that the highest-paying alternative to running your own store would be to manage a branch of a chain store for $49,000 a year. The $49,000 figure, then, is your best estimate of the value of your time. Regardless of what salary you actually assign to yourself, the salary figure entered into the costs of operating your store should be $49,000.

The same procedure can be used for all other scarce inputs that are not bought or hired. Suppose that the best alternative investment of your money would have yielded a return of 15 percent. The cost of investing $20,000 in your business, then, is $20,000 × 15 percent or $3,000 in foregone interest.

Since economic cost includes both accounting costs and an estimate of the value of all inputs that are not included in accounting costs, it comes much closer to giving the true value of the resources used in the process of production.

Sunk costs Economists also stress ***sunk costs*** as a prime instance where opportunity costs differ from accounting costs. Any fixed cost that has already been incurred is a sunk cost. It is irretrievably gone, even if it is now seen as a mistake.

By definition, *once a cost is fixed, it should be forgotten,* and not be allowed to influence what is done next. The invariable economic maxim is: *Bygones are bygones.* Always ignore sunk costs. In plain English: Don't throw good money after bad. Look ahead in your economic decisions, not backward. The box headed "Calculations of Opportunity Cost" illustrates several cases of sunk costs.

Calculations of Opportunity Cost

A. Examples of Commercial Decisions

Apex Products, Inc., bought a grinding mill in 1950 for $10 million, and it has depreciated the full value of the mill since then. Yet the mill is in good shape and could be sold for $15 million at current prices. Apex's use of the mill has an accounting cost of zero. The *opportunity cost* to Apex of using the mill, rather than selling it, is $15 million.

A firm keeps a cash balance of $2 million to meet unexpected difficulties that may arise. It could get a 10 percent return on bonds. The *opportunity cost* of the emergency fund is $200,000 per year ($2 million times 10 percent). The accounting cost of keeping the funds is zero.

A chemical factory releases toxic wastes into a nearby river. This method of waste disposal does not cost the firm itself anything. The cost of the chemicals that the firm produces—its output, in other words—is $8 per pound. But cities downstream must pay $2 million to purify the water. This works out to $2 per pound of the chemical sold. So the *total opportunity* cost of the chemical is $10 per pound, not just the $8 private cost to the firm.

B. Examples of Sunk Costs

A Bad Course. You work hard at a calculus course for four weeks. But you can't get the knack of all those derivatives, and meanwhile your career interests change toward graphic arts. Should you struggle on in calculus to justify the work you have already put into it? No, the time and effort are a *sunk cost.*

Football Tickets. Iowaska University builds a giant football stadium seating 100,000 fans, but then its football program flops and only 20,000 people come to each game. Should it charge them $15 for each ticket, enough to cover all costs, including the investment cost of the stadium, or $3 to cover just the operating costs? The correct answer is $3, for most of the investment cost for the now-too-large stadium is *sunk cost.*

Value of a Car. Mervin Piccolo buys a new sedan for $9,800, but his wife declares it unsafe and tells him to sell it. He tries to sell it for $9,750, but $7,300 is the highest offer he gets. The $2,500 drop in value (from $9,800 to $7,300) occurred when he drove from the dealer's lot onto the street, changing the new car into a used car. The $2,500 is a *sunk cost.*

Economic profit

Economists define profit just as anyone else would: Profit = Revenue minus Cost. Given this definition, most people think of profits as the money left over after all the bills are paid. They are really thinking of revenue minus only direct costs. When costs are defined as economic costs, then

profit has a different meaning. Since the cost figure is calculated on the basis of foregone alternatives, the profit figure also depends on the value of alternatives. ***Economic profit*** doesn't just indicate how much money is left over after all the bills are paid. It tells you how a firm is doing, compared to the best alternative use of its resources. Consider the three possible profit situations:

Economic profits equal to zero If economic profit equals zero, revenue must equal direct plus imputed costs. The firm must be covering its dollar costs and making a return equal to the return it could earn in the highest-paying alternative use of its resources. Knowing this, the firm has no incentive to transfer its resources to another line of production, since it is already making as much as it could in its highest-paying alternative. If a firm's economic profits equal zero, it is making a *normal return*, a return just high enough to keep it in its present line of production. A normal return does not imply that a firm will not try to do better. It simply implies that the firm recognizes that, given its resources and present market conditions, it cannot do better by transferring resources.

Economic profits equal to zero may strike you as undesirable. You are used to thinking of profits as the revenue remaining after the bills are paid. But economic profits are calculated net of *both* the amount needed to pay the bills (direct costs) *and* the amount that could be earned on the firm's own resources (imputed costs). In other words, the accountant's concept of profits includes several imputed costs, whereas the economist's definition of profits includes both the direct costs recognized by accountants and all the imputed costs. To the economist, profit is a return over and above all the costs that must be covered to keep the firm operating indefinitely.

Economic profits less than zero If economic profits are less than zero, revenue is less than the sum of direct and imputed costs. The firm *may* be able to pay its bills, if revenue is at least as great as direct costs. But the firm's return is obviously lower than it would be if it were making its best alternative use of resources. In this case, the firm is not making enough to keep it in its present line of production; it could do better by transferring its resources.

Economic profits greater than zero If economic profits are greater than zero, revenue must be greater than the sum of direct and imputed costs. The firm, then, must be earning a higher return than it could in the highest-paying alternative use of its resources. By transferring its resources to another use, the firm could only do worse. It will therefore stay in its present line of production.

You can see that economic profit is a much deeper concept than simply the dollars left over at the end of the period. Such an economic calculation of profits helps the firm to decide if its resources are being put to the best or most profitable use. In other words, economic profit is the criterion the firm should use in deciding whether to continue in its present business or to transfer its resources to another use.

To illustrate, Jerry and Mary Molnar quit their office jobs as accountants, each at $14,000 per year, for the excitement and independence of running their own sporting goods store, the Fast Track. They both work ten-hour days, seven days a week. After three years, the shop has a steady clientele, $300,000 in sales revenue, and $280,000 in operating costs (including nominal salaries of $7,000 each drawn by Mary and Jerry). They pay $15,000 in interest on the shop's loans and reinvest the remaining $5,000. Their account shows a gain of $5,000. At least they are making some profit.

Or are they? The Molnars could have earned at least $28,000 per year in regular daytime jobs, plus perhaps another $10,000 in spare-time tax accounting (up to their present 70 hours of work per week). The true cost of their labor is therefore $38,000, not $14,000. Because of the $24,000 undervaluation of their own time, they are losing $19,000, not gaining $5,000. Seen another way, they are paying $19,000 a year to have the excitement and independence of running their own business.

Understanding the principles by which a firm should calculate its costs and profits is only one step, although an important one, toward understanding a firm's costs. Now that you have a clearer understanding of exactly what costs are and how they are calculated, it is time to examine the behavior of costs.

To calculate a firm's costs, you really need two pieces of information. First, you must know how much of each input the firm needs to buy to produce its output. Second, you must know what prices it must pay for the inputs it buys, and what prices it should impute to the inputs it owns. By multiplying the input quantities by input prices, you can determine a firm's costs. But how do you know how many inputs the firm will need to buy to produce its output? Clearly, to understand costs, you must first understand the relation between input and output.

Productivity and costs in the short run

Short run and long run

In discussing productivity and cost, economists distinguish between the *short run* and the *long run* because different forces influence productivity, and therefore costs, in the short and long run.

The ***short run*** is a period of time sufficiently brief so that at least one input cannot be varied in quantity. In the short run, there are both *variable inputs*, whose quantity can be altered, and fixed inputs, whose quantity cannot. For some firms, the short run may be very short indeed: an hour, day, or month. For other firms, such as those in the lumber or mining industry, the short run may last for many years.

The ***long run*** is enough time, so that all inputs can be varied in quantity. There are no fixed inputs in the long run.

Labor and raw materials are often used as variable inputs in examples, since their quantities can often be fairly easily adjusted. Capital is used to represent the fixed input, since the quantity of buildings and specialized machinery usually takes the longest to adjust. Throughout this chapter, discussions of production and cost are presented in examples using only two inputs. Once you are comfortable with two-input analysis, you will be able to branch out to more complicated situations involving several variables. (And many of you may find the two-input examples challenging enough.) In the short-run analysis presented next, labor will represent the variable input and capital the fixed input.

Productivity in the short run

Productivity simply refers to how much output a firm can get from its inputs. The more productive inputs are, the more output they can produce. Understanding how output changes when inputs are varied, which means understanding how productivity of inputs varies, is crucial for understanding how costs change as output is varied. The reason is that costs of production are largely determined by the number of inputs a firm needs to buy to produce output.

Output or *product* (the terms are used interchangeably) can be viewed in different ways:

Total product (TP) refers to total output per unit of time (hour, day, week, year, etc.). It is also called "quantity of output" and "*Q*."

Average product (AP) is an output/input ratio, measuring output per unit of some input. In condensed form, Average product = $\frac{\text{Total product}}{\text{Quantity of input}}$. Suppose that *TP* = 50, while quantity of input = 10. The resulting AP of 50/10 = 5 would mean that, on average, 1 unit of input will produce 5 units of output. Another way of stating this output/input ratio of 5 would be that, *on average*, 1 unit of output is produced with 1/5 unit of input. Either way of viewing AP involves the same output-input ratio of 5/1.

Marginal product (MP) measures the change in total product resulting from a one-unit change in input. In condensed form, Marginal product = $\frac{\Delta \text{ Total product}}{\Delta \text{ Quantity of input}}$, where Δ stands for a small specific change in amount. An MP of 3 would mean that an additional 3 units of output could be produced by adding 1 unit of input. Another way of viewing an MP of 3 would be that an additional unit of output could be produced with an additional 1/3 unit of input. Either way of viewing MP involves the same Δ output/Δ input ratio of 3/1.

Total, average, and marginal products are just different ways of viewing the same output/input relations. After you work your way through Table 1 and Figure 1, the relationships among these three product measures should be fairly clear to you.

Start with Table 1. The first two columns indicate how many units of the two inputs are used. Column 1 shows that capital is being held constant at 10 units. You know at once that this is a short-run analysis, since there is a constant or fixed input. Labor is the variable input, as column 2 shows, changing from 0 to 10 units. Given the amounts of inputs used, you can calculate total product or output if you know the *production function*.

The ***production function*** is the underlying technology, which relates inputs to

Table 1 Calculations of total, average, and marginal product

1 Quantity of Capital	2 Quantity of Labor	3 Total Product	4 Average Product	5 Marginal Product
(Q_k)	(Q_l)	(*TP*, *Q*)	(TP/Q_l)	($\Delta TP/\Delta Q_l$)
10	0	0	0	9
10	1	9	9	23
10	2	32	16	31
10	3	63	21	33
10	4	96	24	31
10	5	125	25	19
10	6	144	24	3
10	7	147	21	−19
10	8	128	16	−47
10	9	81	9	−81
10	10	0	0	

Short-run production function:
$y = 10x^2 - x^3$ where y = output and x = quantity of labor.

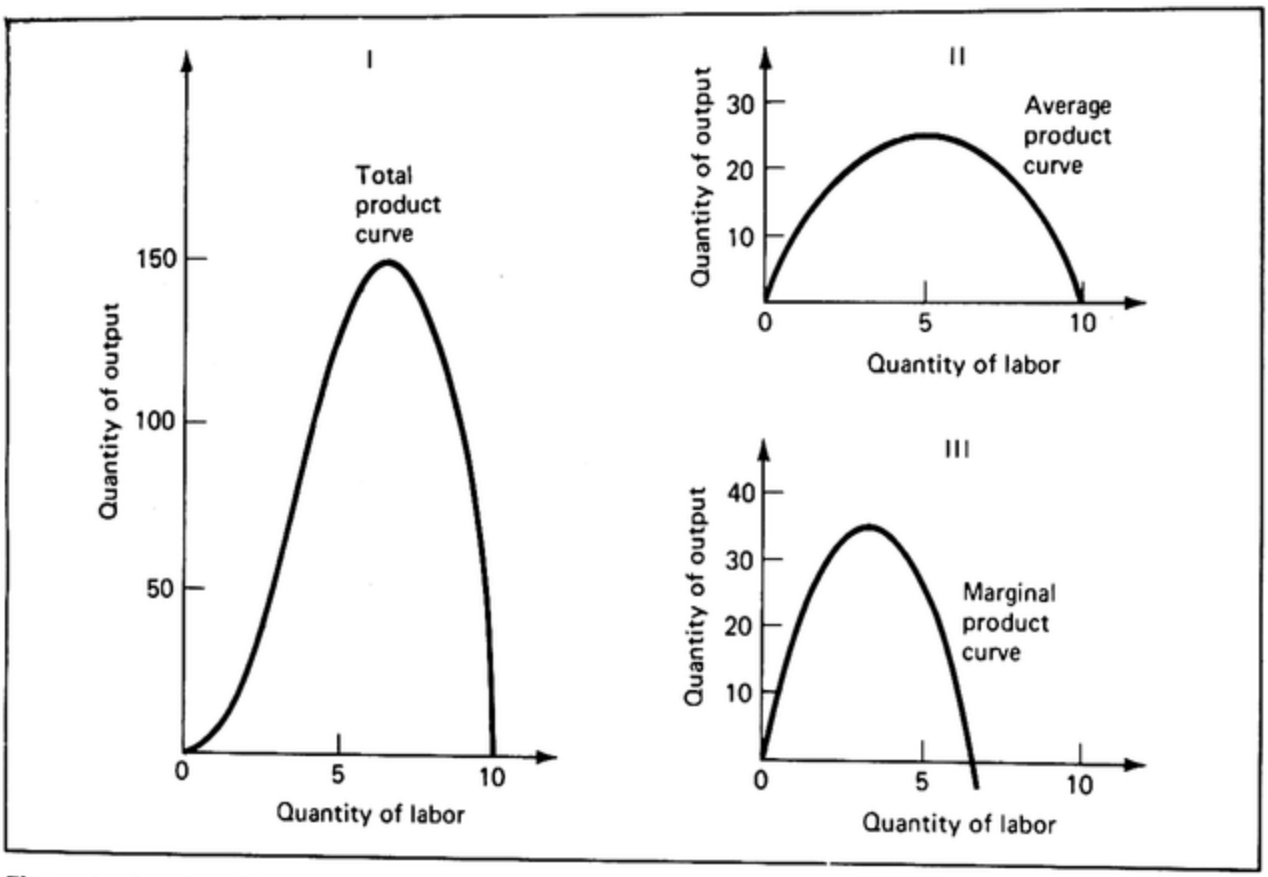


Figure 1 Graphs of total, average, and marginal products from Table 1

output. It is often stated as a mathematical equation. Here you are given a production function of the form $y = 10x^2 - x^3$, where y is total product or output and x stands for the variable input labor. The production function here assumes that capital is held constant at 10 units. If the amount of capital were changed, the production function, and therefore all of the total, average, and marginal figures shown in the table, would have to be recalculated. With the production function and input quantities, you should be able to calculate the total product that would result from the use of different amounts of labor. Try your hand (or, better yet, your hand calculator) at computing total product, and check your answers against those in the table. The total product figures are graphed in Figure 1, Panel I. You can see from the numbers and the graph that total product rises, reaches a maximum, and begins to decline as more labor is added to the fixed amount of capital.

Once you have calculated total product, average and marginal product are easy to arrive at. All the information you really need are total product and quantity of inputs. Dividing total product by the quantity of the variable input will give you the output/input ratio, or average product. This is calculated in column 4 of Table 1 and graphed in Panel II of Figure 1. Average product will rise, reach a maximum, and then decline as more labor is added to the fixed amount of capital. Marginal product is the change in total product resulting from each one-unit increase in the amount of labor, as shown in column 5 of Table 1 and graphed in Panel III of Figure 1. Just like the average product curve, marginal product will increase, reach a

maximum, and then decrease as more labor is used relative to capital.

Once you have grasped the behavior of each individual product curve, you are ready to examine the relations among them. These show up in the relations among the numbers in columns 3, 4, and 5 in Table 1, and are shown graphically in Figure 2.

First think about the relation between TP and MP. The slope of any curve is the change in the vertical/change in the horizontal. The slope of the TP curve must, therefore, be (Δ in output)/(Δ in input). That is the definition of MP. The relation between TP and MP is thus simple: MP is the slope of the total product curve. As long as MP is positive, total product will rise. If marginal product is positive but falling, total product will still increase, but at a slower rate. **When marginal product is zero—as at Point *A*—total product will be at a maximum; and when marginal product is negative, total product will fall.**

All of this may seem obvious to you, but be cautious. Most people are used to thinking about totals and averages, but not marginal figures. People are simply not used to thinking "at the margin," or in terms of small changes. Many of you may be uncomfortable with the idea that total product can rise even though marginal product is falling. That is because you are saying marginal, but thinking total. Remember that marginal product measures the rate of change of output. As long as that change is positive, total product must rise.

Average product (AP) is graphically different from marginal product. It is the slope of a straight line or ray from the origin to a point on TP, as at Point *A* in Figure 2. The ray's slope is the total output divided by the quantity of input, which gives 147/7 = 21 at that point. That slope contrasts with the zero slope of the TP curve itself at Point *A*. Note that the steep-

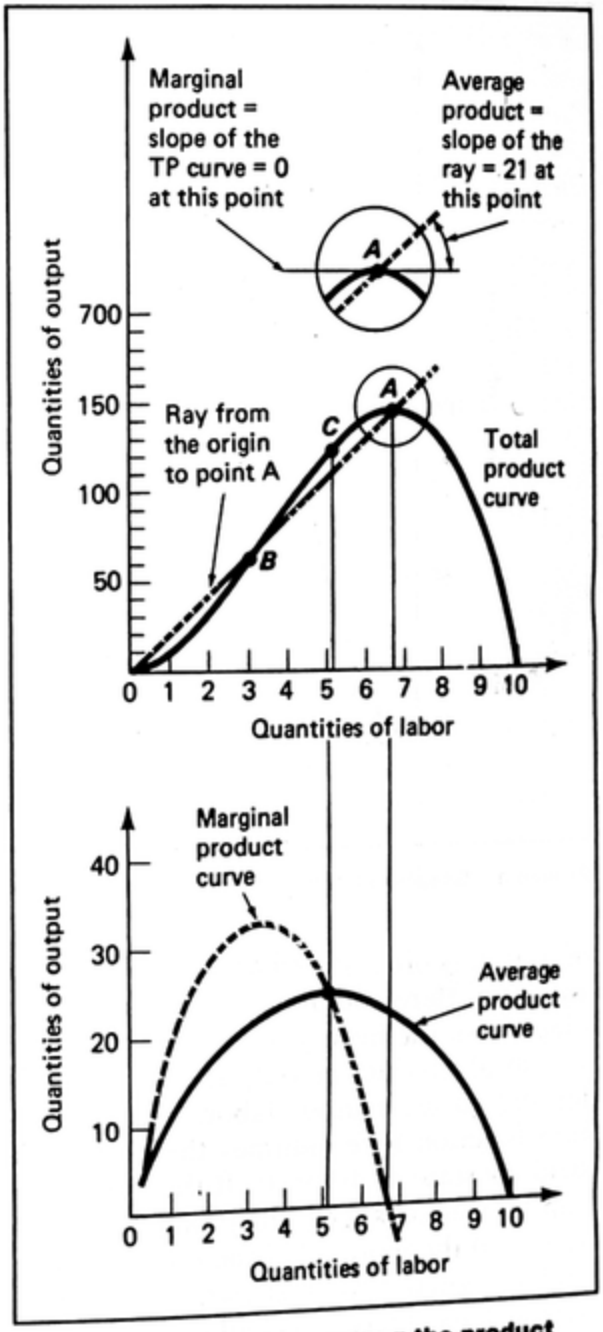


Figure 2 Relationships among the product curves

Total product and marginal product: Marginal product (Δ*Q* output/Δ*Q* input) is the slope of the TP curve. As long as MP is positive, TP will rise. Where MP is 0, TP is at a maximum. This is shown by Point *A*. Where MP is negative, TP is falling. MP will reach a maximum (Point *B*) where the slope of the TP curve is steepest.

Total product and average product: AP is at its maximum at the point where a straight line drawn from the origin is tangent to the TP curve (Point *C*).

Average product and marginal product: If AP is rising, the MP must be above it. If AP is falling, MP must be below it. If AP = MP, AP is at a maximum.

est ray to TP would be at Point *B*. That is therefore the maximum value of AP, as shown in the graph below. The two different slopes—of tangents and rays—embody the conceptual difference between marginal and average product.

Marginals and averages The relation between the marginal and average product curves is especially interesting. The relation between any marginal and average figures has certain simple mathematical properties. As an illustration, think of the one marginal-average relation with which you are probably all too familiar: Your grade point average or GPA and your marginal or additional grades. If you take one course and the marginal or additional grade is above your average, your average rises. If you take another course and the marginal or additional grade for that course is lower than your average, your average falls.

This marginal-average relation is clearly seen by comparing the figures in columns 4 and 5 of Table 1, or by examining the graph of these figures in Figure 2. If AP is rising, the MP must be above AP, pulling it up. The MP itself may be rising or falling. What is important is that it is above the average product. If AP is falling, the MP must be below it, pulling it down. What is significant about the point where $MP = AP$? Up to that point, MP was above AP, so AP was rising. After that point, MP falls below AP, so AP declines. If AP rises up to the $AP = MP$ point and declines thereafter, then the point where $AP = MP$ must be the point where AP is a maximum.

The law of diminishing marginal returns Now that you have had a chance to examine total, average, and marginal products both in numbers and in graphs based on these numbers, the meaning of these product concepts and the relationships among them should be a little clearer to you. But by this time, many of you probably want to know if the numbers in Table 1 were chosen with any purpose in mind. Must AP and MP always rise and then fall? Or could MP and AP first decline and then increase? In other words, is there any reason for the total, average, and marginal curves to be drawn the way they are? As you doubtless suspect, the numbers in Table 1 were carefully chosen to illustrate the expected shape of the product curves. The explanation for these shapes comes from the *law of diminishing marginal returns.*

The law of diminishing marginal returns states that as the quantity of one input is increased, while the quantity of another input is held constant, a point will be reached beyond which additional (marginal) units of input will add less and less to output. In other words, with one fixed and one variable input, a point will be reached beyond which MP must decline, eventually falling below the average. The marginal product curve and therefore the average product curve must have a declining portion. They may increase at first and then decline, as shown in Figure 1, or they may begin to decline from the start, but they must decline at some point.

The law of diminishing marginal returns refers specifically to a situation in which one input is held constant. Therefore, the law of diminishing marginal returns can refer only to *short-run* analysis.

Why does the principle of diminishing marginal returns operate? The answer comes from the changing proportion of inputs. Suppose capital is constant. Now the firm begins to add units of labor. At first, by adding more of the variable input, it finds that it can organize inputs more efficiently, through division of labor or specialization. Each additional unit of input, then, causes output to increase by more than the previous unit. Past some point, however, there may be no more opportunities for specialization. Each unit of labor

Table 2 Effects of additional workers on total, average, and marginal number of meals served

Number of Workers on	Number of Meals That Can Be Served Each Evening	Number of Meals per Worker	Increase in Number of Meals from the Addition of One More Worker
Q Labor	Total Product	Average Product	Marginal Product
0	—	—	
			20
1	20	20.0	
			15
2	35	17.5	
			10
3	45	15.0	
			5
4	50	12.5	

also has less and less capital with which to work. Total product will still increase, but it will increase more slowly with each additional unit of input adding less to output than the units before it.

Consider, for example, the small restaurant that was the case study in the last chapter. Once it is operating, the fixed inputs would be the space bought or rented and the built-in equipment, such as stoves or ovens. The variable inputs would be food, supplies, and labor. Now suppose that when you first open your restaurant you are the only worker. You have to wait on tables, buy food, cook, wash dishes, clean up, figure out bills, and work the cash register. You find that you can serve 20 customers a night.

You decide to hire some help. Now that there are two of you, you can divide up tasks. You do the cooking and cleaning up, while your help waits on tables and works the cash register. With two of you working, you can serve 35 meals a night. (Try calculating the avérage and marginal product associated with the hiring of the second worker, and check your answers against the figures in Table 2.) A third worker allows you to specialize even further. You concentrate on cooking, another worker clears tables and cleans, while the third worker waits on tables and works the cash register. All of you, by organizing more efficiently, can serve 45 meals a night.

A fourth worker allows even more specialization, and you can now handle 50 meals a night. Notice, though, that while the total meals you can serve has increased, the addition of the fourth worker caused output to rise by less than the addition of previous workers had. Output gains from adding workers are still there, but they are smaller. Should the fourth worker have been added? You really can't tell by looking at marginal product alone. You would need to weigh the increase in cost from hiring the worker against the increase in revenue from serving the additional meals. At any rate, the returns from adding workers clearly diminish in this example.

By now, you know a great deal about productivity in the short run. You have been introduced to three different ways of viewing the relations between output and inputs: total, average, and marginal product. You have seen how to calculate these products, how they behave, and why they behave the way they do. Cost curves may seem to have been left far behind. But as you will quickly see, understanding product curves is absolutely essential to understanding cost curves.

Costs in the short run

Since short-run costs are so closely related to short-run productivity, it is not surprising that for each measure of short-run pro-

ductivity there is a cost counterpart. Just as there are fixed and variable inputs, there are fixed and variable costs. Just as there are total, average, and marginal measures of productivity, there are total, average, and marginal measures of costs. In this section, the cost measures are defined, and their relation to productivity is spelled out.

The distinction between fixed and variable costs parallels the distinction between fixed and variable inputs. ***Fixed costs*** **are the costs associated with fixed inputs, which do not vary with the level of output.** Examples of fixed costs would be rent or payments on bank loans. If a firm decides to shut down and produce nothing, its rent and loan obligations will continue for the life of the lease or the loan agreement. These fixed costs must still be paid even if output drops to zero. Besides the fixed costs that must be paid out of pocket, there are imputed fixed costs of using the firm's resources in the current line of production rather than in an alternative. ***Variable costs*** **vary with the level of output.** You can probably think of many examples of variable costs, such as wages for production-line workers and payments for energy and raw materials. These costs vary with output, and if output drops to zero, so will the variable costs.

Besides the distinction between fixed and variable costs, you can also divide costs into categories that parallel the productivity measures: total, average, and marginal cost. As you work through the short-run cost concepts, check your understanding of them by examining Table 3. The first five columns of the table, giving input quantities and total, average, and marginal products, are reproduced from Table 1. Given this information, along with the price per unit of the fixed and variable input, all of the short-run costs can be calculated. The cost figures are graphed in Figure 3.

Total cost (TC) refers to the total cost of production. It is the sum of both fixed and variable costs:

Table 3 ***Calculation of short-run product and cost figures***

1	2	3	4	5	6	7	8	9	10	11	12
Quantity of Capital (Q_k)	Quantity of Labor (Q_L)	Total Product (*TP*)	Average Product (TP/Q_L)	Marginal Product ($\Delta TP/\Delta Q_L$)	Total Fixed Cost (*TFC*)	Total Variable Cost (*TVC*)	Total Cost ($TC = TFC + TVC$)	Average Fixed Cost ($AFC = TFC/TP$)	Average Variable Cost ($AVC = TVC/TP$)	Average Total Cost ($ATC = TC/TP$)	Marginal Cost ($\Delta TC/MP$)
10	0	0	0		50	0	50				
10	1	9	9	9	50	20	70	5.55	2.22	7.77	2.22
10	2	32	16	23	50	40	90	1.56	1.25	2.28	.86
10	3	63	21	31	50	60	110	.79	.95	1.75	.64
10	4	96	24	33	50	80	130	.52	.83	1.35	.60
10	5	125	25	29	50	100	150	.40	.80	1.20	.68
10	6	144	24	19	50	120	170	.35	.83	1.18	1.05
10	6.67	148	22.1	4	50	133.4	183.4	.34	.90	1.24	5.00
10	7	147	21	−1	50						
10	8	128	16	−19	50						
10	9	81	9	−47	50						
10	10	0	0	−81	50						

Note: Maximum output is reached at $Q_L = 6.67$. Costs are calculated with price of capital = 10, price of labor = 20.

Total cost = Total fixed cost
+ Total variable cost
$TC = TFC + TVC$.

Total fixed cost (TFC) does not vary with output. It is equal to the number of units of the fixed input used multiplied by the price per unit of the fixed input. Since TFC is a constant number, it is graphed as a horizontal line.

Total variable cost (TVC) equals the number of units of the variable input used multiplied by the price per unit of the variable input. TVC increases as output increases, with the rate of increase depending on the productivity of the inputs. When inputs are increasingly productive and total product is rising most rapidly, TVC will be rising relatively slowly. This makes sense because when inputs are increasingly productive, increases in output can be sustained with smaller additional purchases of inputs, and therefore smaller increases in costs.

Columns 6, 7, and 8 in Table 3 show the calculations of TFC, TVC, and TC. Working with the data on input quantities and input prices, try to calculate the total costs yourself, and check your answers against those in the table. The total cost figures from the table are graphed in Panel I of Figure 3. If the relationship between output and total cost is well understood, it is also easier to grasp the relationship between output and both average and marginal costs. ***Average total cost*** (ATC) is simply total cost divided by output, or:

$$\text{Average total cost} = \frac{\text{Total cost}}{\text{Quantity of output}}.$$

But to see how average total cost changes as output changes, it is helpful right from the start to express it as the sum of average fixed cost and average variable cost, or $ATC = AFC + AVC$:

$$\text{Average fixed cost} = \frac{\text{Total fixed cost}}{\text{Quantity of output}}.$$

Since total fixed cost is a constant number, as the quantity of output increases, fixed cost *per unit* of output, or average fixed cost, must fall. To check this, examine column 9 of Table 3 and Panel II of Figure 3.

Average variable cost is total variable cost per unit of output, or:

$$\text{Average variable cost} = \frac{\text{Total variable cost}}{\text{Quantity of output}}.$$

Since total variable cost equals the number of units of the variable input purchases multiplied by the cost for each unit of input, the expression for average variable cost can be written as:

$$\text{Average variable cost} = \frac{\text{Quantity of the variable input} \times \text{Price per unit of the variable input}}{\text{Quantity of output}}.$$

Forget for a minute about the input price in this expression and concentrate on the remaining quantity expressions: $\frac{\text{Quantity of variable input}}{\text{Quantity of output}}$. Since this ratio is just the average product of the input turned upside down, the expression for AVC can be written as:

$$\text{Average variable cost} = \frac{\text{Price per unit of variable input}}{\text{Average product of the variable input}}.$$

Now you can clearly see that AVC depends on both price per unit of the variable input and on the average product of the variable input. Specifically, AVC will vary directly with the price of the input and inversely with the average product. This makes perfectly good sense. Obviously, higher input prices mean higher production costs. As input prices rise, the whole AVC curve shifts up, and as input prices fall, the AVC curve shifts down.

Higher productivity should also mean lower cost. Average product, after all, indicates how much input is needed per unit

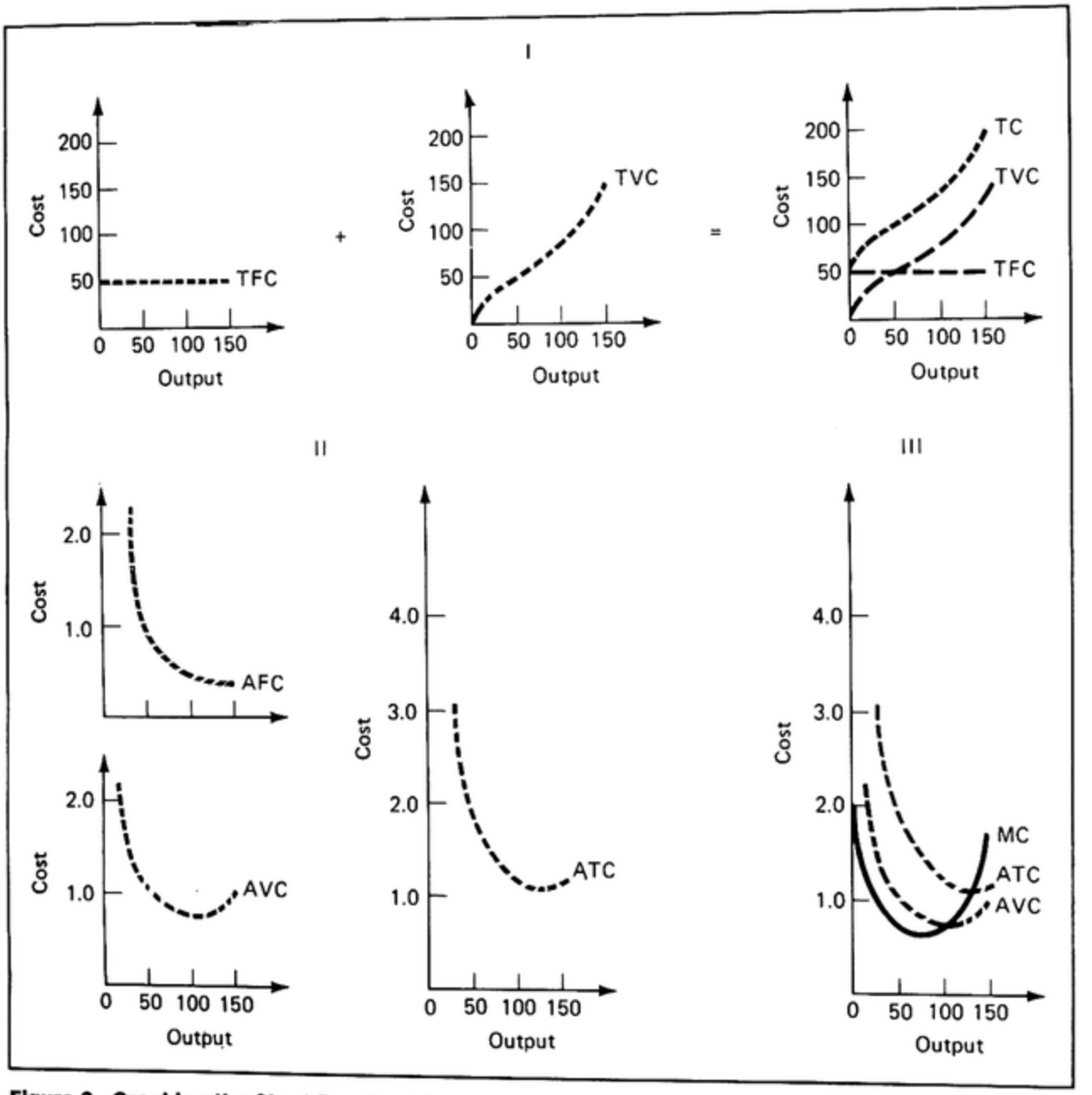


Figure 3 Graphing the Short-Run Cost Curves

Panel I: Total fixed cost is constant, regardless of output. (L) Total variable cost increases with output, at a rate that reflects the changing productivity of input (center). Total cost is the vertical sum of fixed and variable costs (R). Panel II: Average fixed cost falls as output increases (top). Average variable cost falls, then rises, as output rises (bottom). Average total cost is the sum of AFC and ATC. It also falls and then rises (R).

of output. As average product rises, less input is needed per unit of output. Since less input is being bought per unit of output, average costs will be lower. A decrease in AP indicates that more input is needed per unit of output, and since more input is being bought, average variable costs will rise.

The behavior of AVC, then, really depends on the behavior of AP. You already know from the law of diminishing marginal returns that AP may rise and reach a maximum, but then must fall. As AP rises, AVC will fall. When AP falls, AVC will rise. And when AP is at a maximum, AVC will be at a minimum. The relation between AP

and AVC should be apparent from comparing columns 4 and 10 in Table 3 and by studying Figure 3.

Panel II of Figure 3 shows how ATC is derived by adding up AVC and AFC. Since AFC gets smaller as the overhead fixed cost is spread over more units of output, the gap between ATC and AVC narrows as output goes up.

Marginal Cost Besides looking at cost per unit of output, it can also be useful to think about the change in cost from producing an additional unit of output. This measure is called ***marginal cost.*** Marginal cost is concerned only with variable cost, since total fixed costs do not vary with output:

$$\text{Marginal cost} = \frac{\Delta \text{ Total variable cost}}{\Delta \text{ Quantity of output}}.$$

Since the change in total variable cost equals change in quantity of input multiplied by the price per unit of input, the expression can be written as:

$$\text{Marginal cost} = \frac{(\Delta \text{ Quantity of input}) \times (\text{Price per unit of input})}{\Delta \text{ Quantity of output}}.$$

Again, ignore the input price for a moment and concentrate on the quantity expressions. The change in the quantity of the variable input divided by the resulting change in output is just marginal product turned upside down. Thus, we can proceed through the following several steps:

$$\text{Marginal cost} = \frac{\Delta Q \text{ input}}{\Delta Q \text{ output}} \times \text{Price per unit of input.}$$

Since

$$\frac{\Delta Q \text{ output}}{\Delta Q \text{ input}} = MP,$$

$$\frac{1}{MP} = \frac{\Delta Q \text{ input}}{\Delta Q \text{ output}}.$$

Then we can insert $\frac{1}{MP}$ in the earlier equation, thus:

$$\text{Marginal cost} = \frac{1}{MP} \times \text{Price per unit of input}$$

$$\text{Marginal cost} = \frac{\text{Price per unit of input}}{MP}.$$

Or, stated more fully:

$$\text{Marginal cost} = \frac{\text{Price per unit of variable input}}{\text{Marginal product of the variable input}}.$$

Notice the parallel between MC and AVC. MC is input price divided by MP. AVC is input price divided by AP. Again, you can see the relation between production and cost. The law of diminishing marginal returns indicates that MP may rise to a maximum, but then must fall. Therefore, MC must fall to a minimum, but then must begin to rise. Remember that this link between productivity and cost makes sense because productivity is the key to how much input a firm needs to buy. For marginal product, it is the input needed for an additional unit of output. For example, if MP is 5, a firm needs to buy 1/5 of a unit of input to produce an additional unit of output. If MP rises to 10, a firm needs only 1/10 of a unit of input to produce an additional unit of output.

The marginal product–marginal cost relation can be seen by comparing columns 5 and 12 of Table 3, and by studying Figure 4. As long as MP is rising, MC is falling. If MP is falling, MC must be rising. And where MP is at a maximum, with an additional unit of output being produced with a minimum amount of input, MC must be at a minimum.

Finally, Panel III of Figure 3 shows the relationships among marginal cost and both AVC and ATC. You have already seen the relation between AP and MP, and the relation between MC and both AVC and

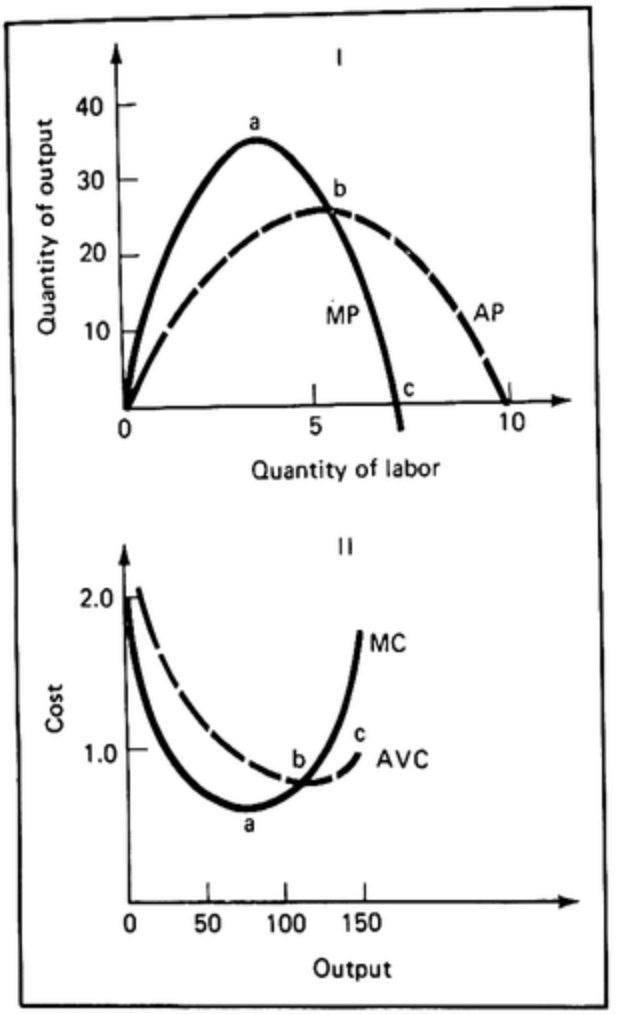


Figure 4 The relationships between product and cost curves

Similarly labeled points correspond to one another. When MP is a maximum (a), MC is a minimum. When MP = AP and AP is a maximum (b), MC = AVC and AVC is a minimum. When MP equals zero (c), output has reached its maximum possible value.

ATC is quite similar. If AVC is falling, MC must be below it, pulling it down. If AVC equals MC, then AVC must be a minimum. If AVC is rising, MC must be above it, pulling it up. The same relation holds for MC and ATC. ATC will decline if MC is below it, rise if MC is above it, and be at a minimum at the point where $MC = ATC$.

This discussion may seem highly abstract. For some students ATC is a theory completely divorced from the real operations of the firm. It is worthwhile to consider for a moment what that ATC curve really represents. Each ATC curve stands for certain combinations of fixed and variable inputs—in other words, a certain technology. The heat and smell and noise associated with actual production are missing, but the results of the use of a particular technology, in terms of cost per unit of output, are effectively conveyed

Now you know how costs behave in the short run and why. The key to cost behavior is productivity: How many units of input are needed to produce output. Knowing how costs behave as output increases is a key to understanding how firms behave. But all of the analysis so far concerns the short run—situations in which there are both fixed and variable inputs. What happens to costs when firms are freed of the constraints of fixed factors—when they are free to choose among any possible combinations of inputs? For the answer to that question, you need to examine the behavior of costs in the long run.

Productivity and costs in the long run

The kinds of decisions that a firm must make in the long run are substantially different from short-run decisions. In the short run, the firm can vary quantity, but only within the constraint of the fixed factor. If a restaurant wants to serve more meals in the short run, the owners can juggle the amounts of labor and food, and, to a certain extent, the numbers of tables and chairs. All of this change, though, must take place within the limits set by the fixed factors, such as the size of the building and the type of built-in cooking equipment.

But if the managers want to serve more customers in the long run, they can vary the amounts of all inputs. They may

lease a larger space or install different amounts and kinds of cooking facilities. In other words, in the long run, a firm is free to choose from among all of the available technologies—from among all of the known ways to combine inputs, none of which is fixed in the long run. By varying the amounts of these factors that are fixed in the short run, it can choose from among all of the existing short-run cost curves.

Derivation of the long-run average total cost curve

The first step for a firm making a long-run decision is to discover the "best" way to produce various output levels. And, of course, the "best" method of production should mean the lowest-cost method of production for a particular output. What the firm needs to develop, then, is a cost schedule or curve showing the lowest cost at which various levels of output can be produced. "Lowest cost" in this context means either lowest TC at each level of output, or lowest ATC at each level of output. The choice of an input combination that will minimize ATC at some output will also minimize TC. It is easier to see what is involved, however, if you focus on ATC.

Such a schedule has to be derived almost a point or section at a time. Select one output level, say 100 units of output. Consider all the possible techniques that could be used to produce this amount of output. Since the short-run ATC curves are shorthand ways of expressing these various technologies or combinations of fixed and variable inputs, graphing these techniques results in a bird's nest of short-run ATC curves, such as Panel I of Figure 5. (For brevity, we will henceforth omit the T for "total" when discussing the short- and long-run cost curves. They will be labeled as SRAC and LRAC.)

In this diagram, one technology or combination of fixed and variable inputs

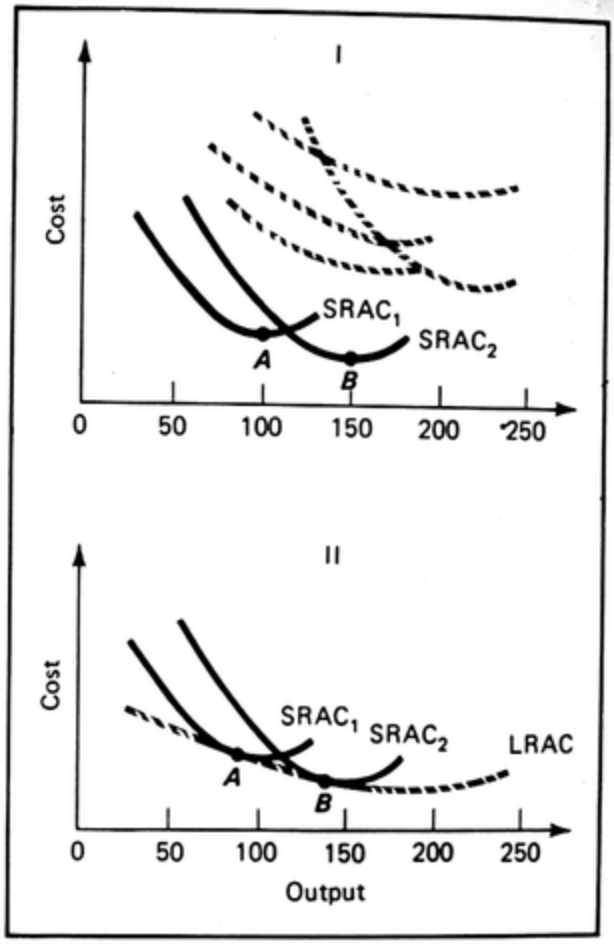


Figure 5 Derivation of the long-run average cost curve

stands out from the others as being the technology that will allow the production of 100 units of output at the lowest per unit cost. This is $SRAC_1$. All other technologies are less efficient in this output range. But you can see from the same diagram that $SRAC_1$ is not the best or least-cost way of producing 150 units. A different technology, using a different amount of the fixed factor, represented by $SRAC_2$, will result in least-cost production.

Identifying the least-cost technology for *all* levels of output involves nothing more than repeated application of the principles for locating the least-cost technology for *each* level. What results is a series of short-run ATC curves, each representing the appropriate technology for a given level of output. Connect all of the points or sections of the short-run curves that represent the most efficient or least-

cost technology for given levels or ranges of output, and you arrive at the long-run ATC curve. This is shown in Panel II of Figure 5. Every point of the long-run ATC curve is tangent to some short-run curve at a point representing the least-cost technology for producing the corresponding level of output.

You can think of the long-run ATC curve as a boundary between what is attainable and what is not. It is what the mathematicians call an envelope, or greatest lower boundary for cost. Since the long-run ATC curve was derived by choosing the least-cost method of production for each level of output, it represents the lowest-cost or economically efficient method of production. Any point below the curve is unattainable, given existing technology and input prices. Any point on or above the curve is attainable. Of course, it is preferable to be operating at a point on the long-run ATC curve, since that represents the lowest costs possible.

The term for operating on the average cost curve is **X-efficiency.** It encompasses all actions that achieve the least-cost outcome. *X-inefficiency* occurs when the firm strays above this least-cost level and incurs unnecessary costs.

Tangencies and minimums The long-run ATC curve is not a difficult concept, but there are some technical points about it that require careful attention. First, the short-run ATC curves are not necessarily tangent to the long-run ATC curve at their minimum points. In fact, a short-run ATC curve is only tangent to the long-run curve at its minimum point where the long-run curve is also at its minimum. This implies that a firm may build a certain-size plant (represented by the short-run ATC curve), planning to use it at a level of output that does not represent the lowest per unit cost possible for that plant.

This may at first seem confusing to you or just plain wrong. The explanation is not difficult, but you have to think carefully about it. First, it would be geometrically impossible for all of the short-run ATC curves to be tangent to the long-run ATC curve at their minimum points unless long-run ATC were constant over the whole range of output and were therefore represented by a horizontal straight line. Remember that when two curves are tangent, their slopes must be equal. Any curve tangent to the downward-sloping portion of the long-run ATC curve must also be downward sloping.

Any curve tangent to the upward-sloping portion of the long-run ATC curve must also be upward sloping. When the long-run ATC curve is at its minimum point, any curve tangent to it must also be at its minimum. The least-cost way of producing a given level of output often involves using a plant either above or below its capacity or least-cost point.

You may accept the geometrical explanation for using a plant at other than its least-cost level of output, but still not understand the economic sense of it. Look at Panel I of Figure 6. Why would the firm produce 100 units with the plan represented by $SRAC_1$? Why not use the plant to produce 150 units at $2 less per unit? This sounds reasonable, but the firm is not interested in using a *given plant* (corresponding to a given SRAC curve) at its most efficient scale. It is interested in *choosing a plant* to minimize the cost of a *given output.*

To see the difference, look at Points *A* and *B* in Panel II of Figure 6. They represent two known ways of producing 150 units of output. You can build Plant 1 and use it at its least-cost point, producing the output at $18 per unit. Or, you can build Plant 2 and use it at an output level that is lower than the least-cost point for that plant. Which plant would you choose? Using Plant 2 to produce 150 units of output

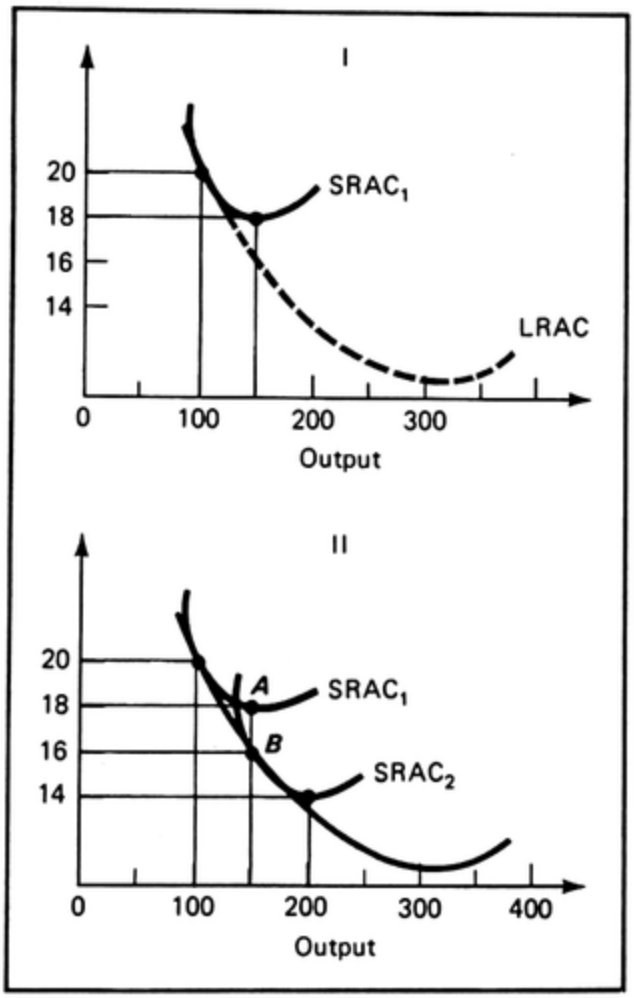


Figure 6 Relationship between short-run and long-run cost curves

seems to involve some inefficiency because you are paying $2 more per unit than if you were using the same plant at its capacity of 200 units.

In fact, however, it is efficient if your goal is to produce 150 units. By building the larger plant, you can produce 150 units at $2 less per unit than you could at the least-cost point of Plant 1. The loss of efficiency ($2) in using the larger plant below its capacity is more than outweighed by the greater efficiency involved in using the larger-scale plant. Remember, a firm's goal is the lowest cost per unit for the amount of output it wants to produce, and not simply to use a plant at the least-cost point.

Choices in the short and long run Now, with the relation between the short-run and the long-run cost curves clearly in mind, you can examine how a producer will make specific choices. Suppose that the producer wants to choose the technology appropriate for an output of 100 units. He or she faces a long-run decision, free to choose any combination of inputs.

The least-cost way of producing 100 units is represented by Point *A* on the long-run ATC curve in Figure 7. That point stands for the technology—the plant size, equipment, amount and type of labor and raw materials—represented by $SRAC_1$. Having decided, the producer then builds the factory, buys the machines, and hires the labor needed for the chosen technology and output. Once the plant is built and the equipment is in place, the producer is back in the short run, with both fixed and variable inputs. As long as the firm continues to produce 100 units, short-run and long-run efficiency coincide.

But suppose that the firm later decides to produce 150 units. Now it is Point *C* that represents the lowest-cost method of production. Since Point *A* and Point *C* are associated with different short-run ATC curves, they stand for technologies using different amounts of the fixed factor. This means that, in the short run, the firm can-

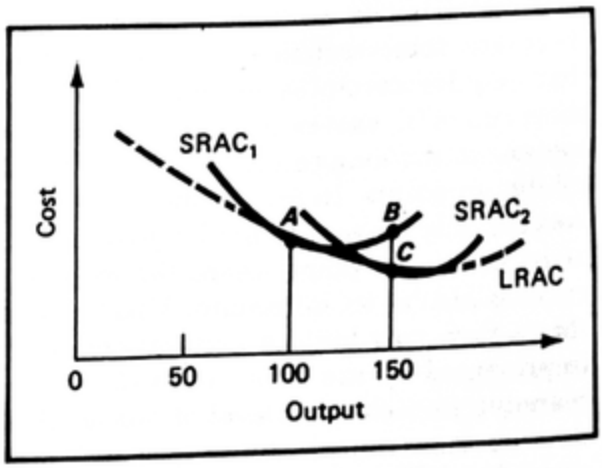


Figure 7 Short- and long-run efficiency

not move from Point *A* to Point *C*. The only way the firm can produce 150 units is represented by a move from Point *A* to Point *B*, changing its variable inputs within the constraints of its fixed inputs. If the firm continues to produce 150 units, it will change the amount of the fixed input as soon as possible, switching from the technology represented by $SRAC_1$ to Point *C* and the technology represented by $SRAC_2$. Movements along the long-run ATC curve can only take place in the long run, when the amounts of all inputs can be varied.

Is it inefficient for the firm to produce 150 units at Point *B* when there is a lower-cost method of production represented by Point *C*? Not in the short run! Given the constraint imposed by the input whose quantity cannot be quickly varied, production at Point *B* is the most efficient way (in fact, the only way) to produce 150 units. In the long run, when the fixed factor can be changed, Point *C* becomes the most efficient way to produce 150 units. So the short-run ATC curves represent short-run efficiency, while the long-run curve represents the points where short-run and long-run efficiency coincide.

You can see that a movement along the long-run ATC curve will occur when the firm chooses a different technology to produce a different level of output at the lowest possible cost. When will the curve shift? The long-run ATC curve is based on two sets of information: the price of inputs and known technology. If either of these changes, the cost curves must shift in easily predictable ways.

If input prices change, the short-run and long-run curves will have to be redrawn. If input prices increase, the curves will shift up, indicating that output can only be produced at a higher price. If input prices fall or if there is a technological breakthrough, so that inputs can be combined in more efficient ways, then average costs fall at each level of output and the long-run ATC curve will shift down. And remember, if the long-run ATC curve is shifting, the short-run curves must also be shifting. After all, if the curves are tangent, one curve cannot shift without the other.

By this time, you have seen how the long-run ATC curve is derived and what it represents. But why is it shaped the way it is?

Economies of scale: The shape of the long-run ATC curve

Since the long-run ATC curve is U-shaped, just like the short-run ATC curves, you might be tempted to say that the shape of the long-run ATC curve must also be due to the law of diminishing marginal returns. Resist that impulse! Remember that the law of diminishing marginal returns refers explicitly to situations in which more of one input is added while another input is held constant. In other words, it can only be used to explain short-run situations.

What about the long-run, then? A different explanation is needed, that of the concept of ***economies and diseconomies of scale.*** (The word *scale* here refers to size.)

The long-run ATC curve represents a variety of combinations of fixed and variable inputs—a variety of technologies. Some of the technologies are best suited for lower levels of output, others for higher levels of output. After all, the best way to build a few specialized cars is not the best way to produce hundreds of thousands of automobiles. To invest $500 million in an assembly plant may be the least-cost technology if the firm is planning to spread that fixed cost over hundreds of thousands of cars. If the firm is planning to sell only a handful of cars, it may want to choose a more labor-intensive technology with lower fixed costs.

Since the long-run ATC curve declines to a minimum, and then begins to rise, adopting larger-scale technology must at

first cause per unit costs to fall. At some point, though, larger-scale technology will cause costs per unit to begin to rise. Why? Why do long-run average costs decline to a point and then start to increase?

Causes of economies of scale The answer lies in the causes of economies and diseconomies of scale. Economies of scale, corresponding to the downward-sloping or declining portion of the long-run ATC curve, have three major causes: specialization, physical laws, and management.

SPECIALIZATION By specializing at a task, by doing it repeatedly, you can become very fast, skillful, and efficient at it. As long ago as 1776, Adam Smith gave ***specialization*** center stage in the Wealth of Nations, using the example of a pin factory:

> One man draws out the wire, another straightens it, a third cuts it, a fourth points it, a fifth grinds it at the top for receiving the head. . . . The important business of making a pin is, in this manner, divided in about eighteen distinct operations. . . . Each person, therefore, making a tenth part of forty-eight thousand pins, might be considered as making four thousand eight hundred pins in a day. But if they had all wrought separately and independently . . . they certainly could not each of them have made twenty, perhaps not one pin in a day. (Adam Smith, *The Wealth of Nations*, 1776 [New York: Modern Library edition, 1937], pp. 4–5.)

The gains in efficiency come from greater dexterity learned through repetition and the saving in time in not having to switch from one task to another. Through specialization, people are sorted into the jobs they do best. By repetition and learning, of course, their skill increases. This is just as valid for today's modern firms as it was for the 18th-century pin factory. The amount of specialization possible and, therefore, the reduction in average costs that results are usually limited by the size of the firm. The larger the plant, the greater the possibility for more complex machinery and a larger labor force, which allows the production process to be broken into an increasingly refined series of specialized tasks. Thus, the larger the plant, the more specialization and the lower the average costs, up to a point.

PHYSICAL LAWS As the scale or size of a plant increases, physical laws—of volume, temperature, motion, metallurgy, and chemical reactions—often help lower costs. A boiler with ten times the volume of another boiler may require only three times as much fuel. A motor's efficiency may become five times greater as its speed doubles. These engineering relations are an important source of scale economies.

MANAGEMENT Larger scale or size may allow the firm to use more advanced and specialized management techniques. Planning and operations may be carried out more precisely. Opportunities available to management in finance, advertising, innovation, and personnel may be better.

Diseconomies of scale and their causes You already know just from the shape of the long-run ATC curve that increases in plant size can only lower average costs up to a point. At some level of output, the possibilities for further specialization through increases in size are exhausted. Expansion beyond that point will not cause any further decreases in average costs and may, in fact, cause *diseconomies* of scale to set in. The reason is simply that the three causes of economies of scale—specialization, physical laws, management—have been carried too far.

SPECIALIZATION can be dull for workers and can lead to resentment and outright rebellion. Short of this extreme, other adverse reactions to specialization can occur.

Workers may grow sloppy; quality may decline; absenteeism may rise. Supervisors may have an increasingly hard time making sure that the work is being done properly. As plant size and specialization increase, all of these reactions may cause costs to increase at some point.

PHYSICAL LAWS too, may result in size hurting, not helping. At some size, a larger boiler may use less fuel but be harder to control or it may crack more often. More complex machinery may be subject to more metal fatigue, friction, or other problems, so that it needs continual repairs. Almost all physical laws reach such negative ranges. When a factory's continued growth causes enough of them to do so, average costs will rise.

MANAGEMENT may turn into an increasingly inefficient bureaucracy as a firm grows larger. Supervisors must themselves be supervised by an additional layer of bureaucracy. More and more layers of administrators and executives are added. Organization becomes more confused, paperwork mounts, and top decision makers may get further and further removed from what is really going on. Good managers can overcome many of these problems. But the tendencies toward trouble are there, increasing as size increases. This is true of all organizations, including private firms.

Typical curves From all of these conditions, economists have come to regard the typical cost curves to be as shown in Figure 8. The average cost curve slopes down, reaches a minimum, and then rises. Marginal cost rises, perhaps steeply, cutting through the minimum of the average cost curve. We will see in the next chapter that marginal cost is the central concept in deriving the supply curve.

Although the LRAC curve declines and then rises, it may also have a flat or horizontal portion, as shown in Panel I of Figure 9. Here the forces affecting cost balance out evenly, so that average costs are constant. Such a portion of the LRAC curve would be a portion of *constant returns to scale.* It might represent a firm increasing output by building identical plants, all at the same average cost level Even here, however, the increasing cost of supervising and coordinating the activities of an ever-increasing number of firms may cause costs of management to rise, eventually bringing on diseconomies of scale.

While all LRAC curves slope down and then up, each industry will have its own specific form of curve, reflecting its own technology. In some industries, the decline of average costs extends up to high levels of output. A firm has to produce a large

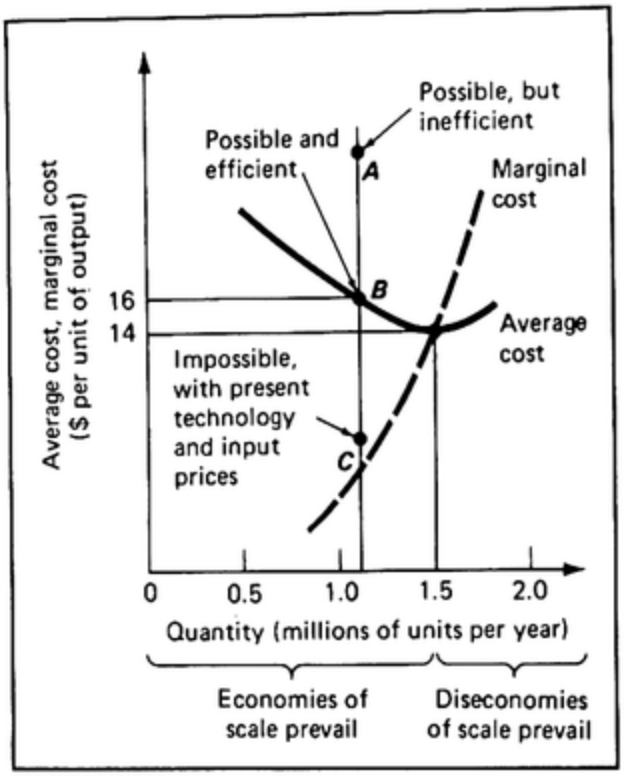


Figure 8 Long-run cost curves for a typical firm: Average and marginal cost

Average cost declines in the range where economies of scale prevail. The firm's natural level of "capacity" is where its average cost is lowest: in this case, 1,500,000 units per year at $14 per unit. At higher levels of production, diseconomies of scale prevail. The marginal cost curve rises sharply above the average cost curve, pulling it up.

Point *A* is possible but inferior to Point *B*. Point *C* is impossible to attain with present technology.

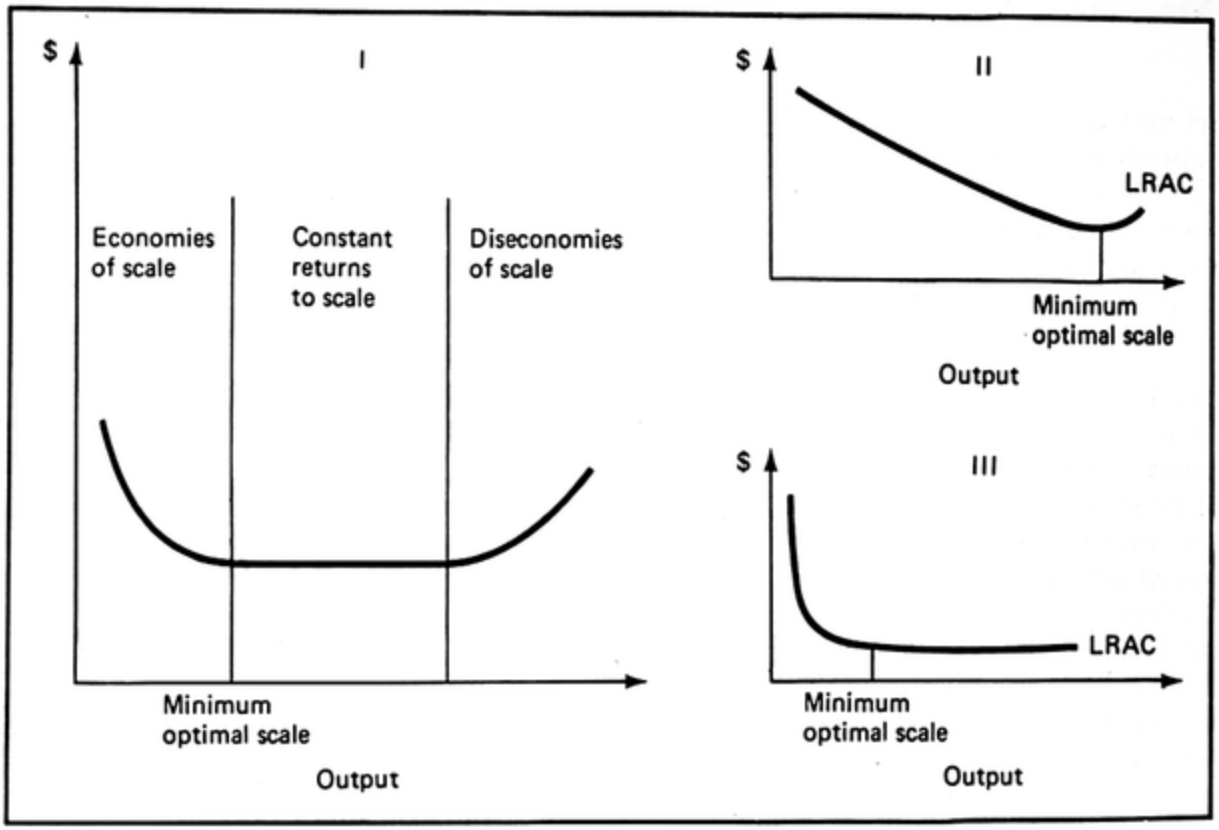


Figure 9 Possible shapes of long-run ATC curves

quantity of output to reach ***minimum optimal scale (MOS):*** the output where average cost is smallest.

Such a case is shown in Panel II of Figure 9. It is called "natural monopoly," because there is room for only one firm that is large enough to reach MOS. An example is electric service in local markets. One electric firm can supply the entire market at a lower cost than could two or more firms, each with higher average costs because its output is below MOS.

The opposite case is "natural competition." It is illustrated in Panel III of Figure 9. Minimum optimal scale is small, so that many efficient firms can coexist and compete with one another within the market. Good examples of that are farming, printing, and retail shops, where most firms are small and have tiny market shares.

Therefore, the shape of the cost curve can be decisive to the chances for vigorous competition in each market. In research since the 1930s, economists have estimated the MOS in over 40 major markets. A selection of the best recent results is given in Table 4. MOS ranges from as little as 1 percent of the market in fabric weaving and shoes up to 16 percent in canned soups.

MOS sets a floor under the actual firms' market shares, because a firm cannot usually survive if its costs are above those prevailing in the market. For competition to be as vigorous as possible, actual market shares would be driven down to MOS. Any market share above MOS is excess, because it gives no improvement in cost. Moreover, it may reduce competition (as we will show in Chapters 10 and 11).

Therefore, the comparison between MOS and actual market shares is important. Table 4 shows that the largest firms do have market shares well above MOS. They also, on average, have a degree of excess market share. Yet, despite these excess market shares, there is a slight correlation between MOS and market shares. Technology has some influence on the size of firms.

Table 4 Optimal size and actual market shares of 17 U.S. Industries, 1965–1970

Industry	(1) Minimum Optimal Scale for the Firm, as a Percent of the Total U.S. Market (1965–1970)	(2) Actual Market Shares, 1970	(3)	(4) Excess Market Share
		Dominant Firm (%)	Top 3 Firms (average) (%)	Leading Firm (%)
	(%)			
Automobile	15	55	28	40
Batteries (storage)	2	27	18	25
Beer	6	19	13	13
Bearings	6	18	14	12
Cement	2	12	7	10
Cigarettes	9	28	23	19
Computers	15	66	26	51
Detergents	12	48	27	36
Fabric weaving	1	12	10	11
Glass bottles	5	29	22	24
Oil refining	5	18	8	13
Paints	2	13	9	11
Photographic film	14	65	27	51
Refrigerators	15	24	21	9
Shoes	1	8	6	7
Soup (canned)	16	75	31	59
Steel	3	22	14	19

Sources: F. M. Scherer, *Industrial Market Structure and Economic Performance*, rev. ed. (Chicago: Rand McNally, 1980), pp. 94–119; L. W. Weiss, in R. T. Masson and P. D. Qualls, eds., *Essays on Industrial Organization in Honor of Joe S. Bain* (Cambridge, Mass.: Ballinger, 1976); W. G. Shepherd, *The Economics of Industrial Organization* (Englewood Cliffs, N.J.: Prentice-Hall, 1979), Chap. 11; and various other industrial sources.

Many of the MOS estimates are hotly debated, for they bear on the standing of the leading firms in those industries. Thus, General Motors (automobiles), Eastman Kodak (photographic film), and IBM (computers) each control over half of their markets (as shown in column 2), and they often say that their dominance merely reflects their economies of scale. Yet, Table 4 suggests that scale economies extend only up to about 15 percent of their markets. This highly sensitive point will come up for more discussion in Chapters 10–12. It can be decisive in antitrust actions to increase competition.

The marginal conditions necessary for least-cost production

Only one issue remains to be dealt with in this chapter. To draw the long-run ATC curve, it is necessary to find the least-cost method of production for all of the relevant levels of output. If this least-cost method of production has indeed been found, then a certain relation between the marginal products and prices of inputs must hold.

To look at a two-input case: If a firm is indeed producing at the lowest possible cost, then $MP_K/\text{Price}_K = MP_L/\text{Price}_L$ where K = capital and L = labor. It is easiest to grasp the sense of the equality if you first concentrate on the meaning of a single ratio: MP/price. Marginal product represents the increase in output from a one-unit increase in input. If MP is 10, then the last unit of input gave a return of 10 units of output. If you divide MP by price, what do you end up with? The MP/price ratio represents the return, in terms of additional output, from the last dollar spent on the input. For example, if $MP_K = 10$, and $\text{Price}_K = \$5$, then the last dollar spent on capital resulted in a two-unit increase in

output, since \$5 bought 10 units of output at the margin.

Suppose that while the $MP_K = 10$, and the $\text{Price}_K = \$5$, then $MP_L = 21$, and $\text{Price}_L = \$3$. The resulting ratios would be:

$$\frac{MP_K}{Price_K} = \frac{10}{5} = 2 \text{ while } \frac{MP_L}{Price_L} = \frac{21}{3} = 7.$$

In this case, $MP_L/\text{Price}_L > MP_K/\text{Price}_K$. The last dollar spent on labor gave a return of 7 units of output, compared to the last dollar spent on capital, which gave a return of only 2 additional units of output. The dollars that the firm is spending on its inputs are not being equally "productive." This means that there is an opportunity for the firm to switch some dollars from capital to labor and to increase output without increasing cost.

In this case, if a firm reduces its spending on capital by \$1, it loses approximately 2 units of output. Spending this \$1 on labor causes an increase of approximately 7 units of output. The switch of \$1, then, resulted in a net gain of 5 units of output. As long as the MP/price ratios are not equal, there is an opportunity for such favorable switching of input dollars.

As more and more dollars are switched from capital to labor, what happens to the MP/price ratios, assuming that the price of inputs remains constant? As dollars are switched from capital to labor, the quantity of labor increases relative to that of capital, and the marginal products of capital and labor begin to change. Each unit of labor has less capital with which to work, and the MP_L begins to decrease. Each remaining unit of capital has more labor with which to work, and its marginal product begins to rise.

As dollars are switched from purchases of labor to purchases of capital, the MP_L/Price_L decreases and the MP_K/Price_K increases. Switching dollars from one input to another should continue until $MP_L/\text{Price}_L = MP_K/\text{Price}_K$. At this point, the last dollars spent on all inputs yield equal returns in terms of additional output. Since there is no further opportunity for favorable changes in input proportions, the least-cost combination of inputs has been achieved.

The MP/price equality can, with a simple regrouping of terms, be written as:

$$\frac{MP_K}{MP_L} = \frac{\text{Price}_K}{\text{Price}_L}.$$

This looks at the same issue in a slightly different light. This expression emphasizes that a firm should adjust its inputs until the ratios of the marginal productivities of inputs equal the ratios of input prices. The sense of this is fairly easy to grasp. Most firms must take input prices as given. If $\text{Price}_K = \$10$, and $\text{Price}_L = \$5$, there may be little a firm can do to change the price ratio. What the firm *can* do is adjust quantities of inputs until the last unit of capital is twice as productive as the last unit of labor. A firm is willing to pay twice as much for capital only if it is twice as productive at the margin.

Understanding the MP/price rule for choosing the most efficient method of production can help to illustrate one of the major advantages of a relatively free market system. Suppose that the firm has achieved the combination of inputs at which $MP_K/\text{Price}_K = MP_L/\text{Price}_L$. Now suppose that labor becomes scarcer relative to demand, so that its price increases. With Price_L increasing, the value of MP_L/Price_L falls, while MP_K/Price_K remains constant. Now, $MP_K/\text{Price}_K > MP_L/\text{Price}_L$. To restore equality, the firm must increase its purchase of capital and reduce its purchases of labor until MP_L/Price_L once again equals MP_K/Price_K.

The increase in the price of the input causes the firm to use less of the input that has become scarcer and therefore rela-

tively more expensive. To the extent that prices really do reflect relative scarcities, the drive of the firm to achieve a balance between input productivity and price will cause increasingly scarce inputs to be used more sparingly.

For example, the price of airplane fuel rose from 11 cents per gallon in 1973 to $1 per gallon in 1980. Since other input prices for airlines rose only about 50 percent, fuel had become relatively much scarcer. The airlines' marginal product/price ratios were now sharply out of line. The airlines responded by making drastic efforts to conserve fuel. That moved them back up their marginal product curves, and *that* would move the ratios between MP and prices back toward equality.

To do this, the airlines switched to lighter seats, rugs, meal trays, and belt buckles. Some stripped the paint off the planes, saving up to 600 pounds per plane. Cargo holds were revamped and wall linings removed. Flight patterns were shifted toward higher altitudes and cruising speeds were reduced. Fuel reserves carried on the planes were cut.

These changes illustrate the wide range of choice—even on planes already built—for responding to input price changes. The same applies to factories and equipment of many kinds. The equal marginal ratios merely show in a formal way the directions that intelligent managers are constantly moving.

Summary

The purpose of this chapter was to introduce you to the concepts of short- and long-run productivity and cost.

1. *Technology* or *the state of the art* encompasses all of the known methods of production.
2. Given technology and input prices, the firm's goal is to choose the method of production that will allow a given level of output to be produced at the lowest possible cost.
3. Economists measure costs using the concept of *foregone alternatives*. Costs calculated according to this concept are called *economic costs* or *opportunity cost*.
4. The opportunity cost of inputs that are bought or hired is calculated by adding up the dollars paid out by the firm. These are the *direct* or *dollar costs* of the firm.
5. *Economic profit* equals revenue minus economic cost.
6. Discussions of productivity and cost must distinguish carefully between the *short run*, a period of time during which at least one input is fixed, and the *long run*, in which all variables can be altered in quantity.
7. Output can be measured in three different ways:
   a. Total product or output.
   b. Average product or output per unit of input.
   c. Marginal product or the change in output from one unit change in input.
8. As the *law of diminishing marginal returns* states, in the short run, as more of the variable input is added to the fixed input, a point is reached beyond which *marginal product* and *average product* must decline.
9. There is an important relation between marginal and average measures. If an average is rising, the marginal must be above it, pulling it up. If an average is falling, the marginal must be below it, pulling it down.
10. For each of the three measures of output, there are corresponding measures of costs:

a. Total cost is the sum of total fixed cost and total variable cost.
b. Average total cost is the cost per unit of output. It is the sum of average fixed cost and average variable cost.
c. Marginal cost is the change in total cost from a one-unit change in output.

11. Short-run costs are determined by the cost per unit of output and productivity, that is, by the amount of input needed to produce total, average, or additional units of output.

The main points concerning *long-run* conditions are:

12. To derive the long-run ATC curve, it is necessary to identify the least-cost method of producing various levels of output. This involves selecting the short-run ATC curve that represents the lowest-cost technology (combination of fixed and variable inputs) for each given output level. When the points or sections of the short-run curves that represent these least-cost technologies are connected, the resulting schedule will be the long-run ATC curve.
13. In the long run, a firm is free to choose any point on the long-run ATC curve, that is, any combination of fixed and variable inputs. Once a point is selected, and the plant and equipment representing that technology are built, the firm is operating in the short run. It can vary its output only within the constraints of its technology. Selection of a new point on the long-run ATC curve can occur only in the long run, when all inputs can again be varied.
14. The long-run ATC curve, and the short-run ATC curves from which it is derived, will shift if input prices or known technological possibilities change.
15. The U-shape of the long-run ATC curve can be explained by economies and diseconomies of scale. As a firm adopts larger-scale technology, average costs will at first drop because of the influence of specialization, physical laws, and management. Expansion beyond a certain point may cause rising average costs because these three influences have been carried too far. A long-run ATC curve may also have a horizontal portion representing constant returns to scale, where forces affecting cost balance out.
16. While all long-run ATC curves decline and then rise, each industry will have a specific shape reflecting its own technology. Average costs may fall off slowly or sharply as output rises.
17. If a firm is producing its output at lowest cost, then the MP/price ratios for all inputs must be equal. If the MP/price ratios are not equal, there is always an opportunity for changes in input proportions that will lower costs and raise profits.

Key concepts

Technology
Opportunity cost
Direct cost, or accounting cost
Imputed or implicit cost
Sunk cost
Economic profit
Short and long run
Total, average, and marginal product
Production function
Total fixed cost

Total variable cost
Average total cost
Average variable cost
Marginal cost
Economies and diseconomies of scale
Specialization
Minimum optimal scale
X-efficiency

Questions for review

1. Two firms both produce identical output. Each firm uses a different technology.
   a. What would be some reasons for the choosing of different technologies?
   b. Is one firm necessarily making a mistake in its choice of technology? Explain.
2. Consider opportunity cost in each of the following situations.
   a. You can either study for an exam or go to a party. You decide to study for the exam. What type of information would you have to know to determine if the opportunity cost of your studying was high or low?
   b. You decide to earn some money this year at school by acting as a go-between for an artist friend of yours and the students on your campus. You will advertise, take orders, pick up the finished items, and deliver them.
      i. Make a list of all the items you would want to include under direct cost and under estimated or imputed cost.
      ii. To determine the price that will give you a fair return for your effort, you need to estimate the opportunity cost of your time. Explain how you would do this, and see if you can come up with a dollar figure representing the value of an hour of your time.
   c. You buy a $20 concert ticket, decide that you really do not want to go, and find that you cannot sell the ticket at any price. What is the opportunity cost of:
      i. going to the concert, although you know that you will dislike it;
      ii. skipping the concert and "wasting" the ticket.
3. Two students are arguing about the concept of economic profit. One claims that to imply that firms should continue to produce with an economic profit of zero is antibusiness. Firms *need* a profit to reinvest in the business and *deserve* a profit as a reward for their effort. The other student claims that, on the contrary, an economic profit of zero means that the firm *has* money for reinvestment and *is* being rewarded for its effort. They ask you to settle the dispute. Do so!
4. Classify the following business decisions as short-run or long-run decisions.
   a. A firm wants to make a seasonal adjustment in output to meet higher Christmas demand.
   b. A group of business people have decided to produce a new product and are attempting to determine the best technology to use.
   c. A firm finds that demand for its product is lower than anticipated, so it is trying to decide the most efficient method of laying off some workers.
   d. Given dramatic increases in fuel prices, a manager must deter-

mine whether the firm's present technology is now outmoded.

5. You own a small business and keep very careful records on inputs, output, and cost. One of your employees, the fourth worker hired, sees that when she was hired, output increased by more than it did when the second, third, or fifth workers were hired. She claims that this proves she is more productive and skilled than the other workers. Show that this may not be true.
6. You find a friend of yours struggling with some of the marginal concepts in the chapter. He presents you with his latest confusion: "If marginal cost, the addition to cost, is declining, then it seems to me that total cost should also be declining. Yet MC can decline, while TC can only increase." Explain clearly (and patiently) to your friend exactly what his confusion is.
7. A long-run ATC curve is derived from tangencies with short-run cost curves. Since the long-run ATC curve represents the least-cost method of production for different levels of output, the tangencies between the long-run and short-run average cost curves must therefore occur at the minimum points on the short-run ATC curves. True or false? Explain.
8. A firm that has adjusted its variable input to meet an unexpectedly high demand for its product finds that it is producing at a point higher than its calculation of long-run ATC for that level of output. Does this necessarily imply inefficient production? Explain.
9. Can the fact that long-run costs decline and then increase be explained by decreases and increases in input prices? Why or why not? If not, what influences *do* explain the behavior of long-run average costs?
10. A firm finds that the addition of its last unit of capital caused output to increase by 15 units, while the addition of its last unit of labor caused output to rise by only 7 units. Since the last unit of capital is obviously more productive than the last unit of labor, can the firm conclude that it should purchase more capital and less labor? Explain.

9

Pricing and Output Under Perfect Competition

As you read and study this chapter, you will learn:

- the simple rule for maximizing profit
- the nature of the firm's supply curve
- the derivation of short-run and long-run market supply curves
- the various conditions reached in efficient production in competitive markets

A pair of scissors requires two blades to cut paper, just as it takes two hands to clap. We are now ready to construct the second blade of the microeconomics scissors, by deriving the market supply curve. Working together with demand, the conditions of supply determine the levels of price and output in competitive markets throughout the economy.

As you saw in the last chapter, supply curves reflect cost conditions. In this chapter, we show how the firm compares those costs with prices to determine its rate of production. The result in a competitive market is a supply curve for the firm, which coincides with the marginal cost curve. Then we demonstrate how those individual supply curves determine the market supply curve. Along the way we also discuss shifts in the curves and the various slopes they may have.

The final section notes the efficient conditions that prices and outputs reach in competitive markets. That is an important point

of reference for the next several chapters because, as we will see, monopoly leads to inefficiency.

The rules for maximizing profits

From the preceding chapter, you know how the firm minimizes its cost of producing any given level of output. Now we move to the firm's final step: ***choosing the output level that maximizes the firm's profits.***

That, in turn, divides into two separate company decisions. The first is *whether the firm should produce at all in this market.* Under some conditions, the firm will close down and shift its resources elsewhere. Second, if the firm does decide to continue to produce, *it must still decide how much to produce.*

Should the firm produce at all?

To answer this basic question, one must use the concept of *opportunity cost.* In the long run, a firm should continue operating in its market only if its total sales revenues are large enough to cover the opportunity costs of the resources its uses. If this occurs, then the firm is covering all of its operating costs and making as high a return on its resources as it could in any alternative use of its resources.

To put the production decision in unit terms, divide total revenue and cost by the quantity of output. A firm will only produce in the long run if:

$$\frac{\text{Total revenue}}{\text{Quantity}} \geq \frac{\text{Total cost}}{\text{Quantity}}$$

$$\text{or Price} \geq \text{Average Total Cost.}$$

What about the short run? In this perspective, resources cannot be easily transferred to another use, since by definition there are some fixed inputs that cannot be varied in quantity. In the short run, cost does not drop to zero even if production ceases. The fixed costs, such as bank loan payments or rent, must still be paid. Therefore, in the short run, it might be most advantageous for a firm to operate at a loss, if by doing so, it can cover its operating costs and pay off at least some of its fixed costs.

How much should the firm produce?

Profit equals revenue − cost. Therefore, ***the rules of profit maximization*** must involve the weighing of costs and revenues. To consider the effect of a one-unit change in output on revenue and cost, we must use marginal concepts. If a firm has reached the profit-maximizing point of operation, then a one-unit increase or decrease in output should reduce its profits. The change in profits that results from a one-unit change in output can be expressed as the change in revenue minus the change in cost, or:

$$\frac{\Delta\text{Profit}}{\Delta\text{Quantity}} = \frac{\Delta\text{Revenue}}{\Delta\text{Quantity}} - \frac{\Delta\text{Cost}}{\Delta\text{Quantity}}.$$

The change in cost resulting from a one-unit change in output is *marginal cost,* a concept introduced in the last chapter. The change in revenue resulting from a one-unit change in quantity is called, as you might expect, ***marginal revenue.*** The change in profit from a one-unit change in output can be expressed as:

$$\text{Marginal profit} = \text{Marginal revenue} - \text{Marginal cost, or } MR - MC.$$

If profits are rising (or losses are falling) from the production of additional units of output, the additional units must be adding more to revenue than they are to cost. The firm must be producing where $MR > MC$. If profits are falling (or losses rising) from the production of additional units of output, the additional units must be adding more to cost than they are to

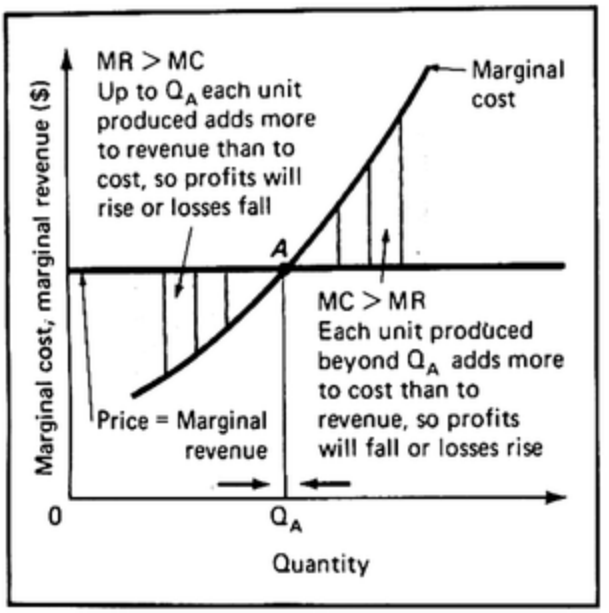


Figure 1 The profit-maximizing level of output

revenue. The firm must be producing where $MC > MR$.

To see this clearly, consider Figure 1, which shows marginal cost and marginal revenue. You already know the shape of the *marginal cost curve*. It first falls and then rises; we will work only with the upsloping portion of this curve. The *marginal revenue curve* may take one of two shapes. It may be horizontal, or it may be downward sloping, depending upon the shape of the firm demand curve. This will be explored in detail later.

For the moment, the MR curve is drawn as a horizontal line. This indicates that the firm always receives the same amount of additional revenue from selling additional units of output. (Either a horizontal or a downward-sloping marginal revenue curve would work equally well to illustrate the point being made.) The marginal revenue and marginal cost curves are combined in Figure 1. The diagram looks fairly simple, but it illustrates one of the most important concepts in microeconomics.

At the level of output marked Q_A, marginal revenue just equals marginal cost: $MR = MC$. To the left of Q_A, $MR > MC$. In this leftward area, the firm will be adding more to revenue than to cost each time it produces an additional unit of output. Therefore, the firm will be increasing its profits (or reducing its losses) if it increases its level of output. As the level of output approaches Q_A, the gap between MR and MC narrows. Each additional unit of output adds less to profit. Nonetheless, as long as $MR > MC$, the firm does increase profits as it produces additional units, and it should continue to increase output up to Q_A. To the right of Q_A, $MC > MR$. Each additional unit produced adds more to cost than to revenue, driving the firm's profits down (or losses up). If the firm finds itself producing in the region to the right of Q_A, it should reduce output back to Q_A.

It is obvious that producing to the left of Q_A is not profit maximizing. The firm can add to its profits by expanding output up to Q_A. Producing to the right of Q_A is not profit maximizing either. The firm can do better by contracting output back to Q_A. The profit-maximizing point, then, must be Q_A, the level of output at which $MR = MC$. To produce one unit less than Q_A means giving up a unit that added to the firm's profits (or reduced its losses) by adding more to revenue than to cost: To produce one unit more means producing a unit that will reduce the firm's profits (or increase its losses) by adding more to cost than to revenue. It is only at Q_A that the firm is doing as well as it can, in the sense that no adjustment in output levels will improve its profit position.

Remember, though, that the relationship between MC and MR will only tell you whether the firm is maximizing its profits. It will not show the level of the firm's total profits. The gap between the marginal revenue and marginal cost curves does not represent profit. It represents only *potential changes* in profits or

losses, with no indication of the overall profit-and-loss level. To determine the firm's absolute level of profits, one must know the average or total cost and revenue measures, and not just marginal measures.

To summarize, what rules must a firm follow to maximize profits? ***In the short run,*** **the firm should produce if** $P \geq AVC$ **at the level of output at which** $MR = MC$. **In the *long run,* the firm should produce if** $P \geq ATC$ **at the level of output at which** $MR = MC$. Large or small, single proprietorship or giant corporation, a firm must follow these rules if it wishes to profit maximize. Although all profit-maximizing firms follow the same rules, the market outcomes differ greatly, depending upon the market in which the firm operates.

Setting output under perfect competition

The nature of pure competition

The ideal model of competition is *pure competition.* It applies to an "atomistic" market containing 50 or more firms, each of which has a negligible share of the market. Each firm is too small to influence the going market price by any action of its own. Therefore, each firm is a *price taker:* It takes the market price as a given, which it cannot change. Accordingly, the firm's own demand curve is a horizontal line at the going market price. By comparing its costs with that going price, the firm can choose the level of its output that will maximize its profits.

Economists use that pure model of competition as the setting for deriving the market supply curve. Yet, real competition in real industries is a richer phenomenon than the ideal case of pure competition. Some students therefore doubt at first that the pure competitive model and its results are generally valid. Nevertheless, the competitive model is actually a good approximation of what competition does accomplish in the main mass of real markets. ***If there is intense rivalry among as few as two firms, the essentials of competitive supply, pricing, and efficiency may still occur.***

To demonstrate that important point, we will first present the substance of competition and then show the basic unity among competitive processes. The competitive assumptions will immediately enable us to derive the supply and efficiency conclusions of this chapter. The concepts also underlie all the rest of microeconomics. Because competition is so important to microeconomics, it needs a careful review at this point to prepare you for the next 12 chapters.

Competition: a process and a zone of choice

Competition is a prime force by which free choice is transmitted into efficient resource allocation.

Effective competition is balanced To be effective, the competitive process needs to be *open and free,* not predetermined, so that effort and skill are needed to compete successfully. The competitors must also be *comparable.* A contest between unequals is not genuine competition. If one competitor has sharp advantages, then the "competition" is not meaningful. The different weight classes in wrestling and boxing recognize this: A bout between a heavyweight and a bantam-weight is punishment, not competition.

Also, true competition is a continuous process. A foot race or a game of Monopoly will end; the competition is over once the prizes are gained. In real life, by contrast, competition in markets means a continuous mutual striving among more-or-less equal firms. If, instead, one competitor gains dominance over the market, then ef-

fective competition is replaced by a degree of monopoly.

Rivalry and pure competition may give similar results.

Rivalry Even if there are only two competitors, their ***rivalry*** may be intense. Though one of the two firms may get the upper hand for a while in a market, the other may soon fight back and equalize its share of the market and profits. Such a rugged rivalry may stir great efforts from the firms and force their prices down close to the levels of their costs. Therefore, effective competition is possible even when there are as few as two or several firms. Each firm would then think and act as if it had little leeway to raise prices or to earn excess profits. If the two firms colluded with each other, they could wield monopoly power. But their intense rivalry prevents any such collusion.

Pure competition, by contrast, guarantees that prices will be forced down to the level of costs. It is the economists' ideal version of competition, an ideal model that has been honed to precision. Like other economists, we will use it to derive the conditions of supply, price, cost, and efficiency.

The analysis of pure competition rests on these four assumptions:

1. There is one identical good sold in the market.
2. No firm has a significant share of the market.
3. All firms adjust rapidly to any changes.
4. There are no hindrances to movement into and within the market. Entry by new producers is free, and changes are frictionless and quick.

For ***perfect competition,*** there is one further assumption: Each firm knows everything about demand and supply conditions in the market.

Perfect competition is an ideal version, designed to show the analytical outcomes as clearly as possible. Does this mean that it differs from "real" competition, such as rivalry? Not necessarily. *All forms and degrees of competition—including rivalry—lead toward the same basic kinds of economic results.* They narrow the firm's range of choice over its price, forcing it to reduce costs and to adopt new techniques. Though strong rivalry may not always keep prices down exactly to minimum cost levels, it will keep them close to those levels. As long as there is some degree of competition, there is some pressure.

Firm and market demand in perfect competition

In perfect competition, any one firm is so small relative to the market that the amount that it sells has no impact on market price. Whether an individual farmer brings 200 or 2,000 bushels of wheat to market will have no impact on the domestic or world price of wheat. Because a firm in perfect competition can sell as much output as it chooses at the prevailing market price, the ***firm demand curve*** *is a horizontal line at the level of market price,* as shown in Panel I of Figure 2. Panel II emphasizes that the market price is determined by the interaction of supply and demand forces in the market.

The marginal revenue curve of a perfectly competitive firm is easy to derive. Marginal revenue is the change in total revenue that results from selling an additional unit of output. Firms in perfect competition can sell as many units of output as they wish at the going market price. Therefore, each time a firm sells an additional unit of output, the addition to revenue or marginal revenue equals the market price. The marginal revenue curve, then, is a horizontal line at the level of market price, coinciding with the firm's demand curve. Total revenue (or price × quantity) increases

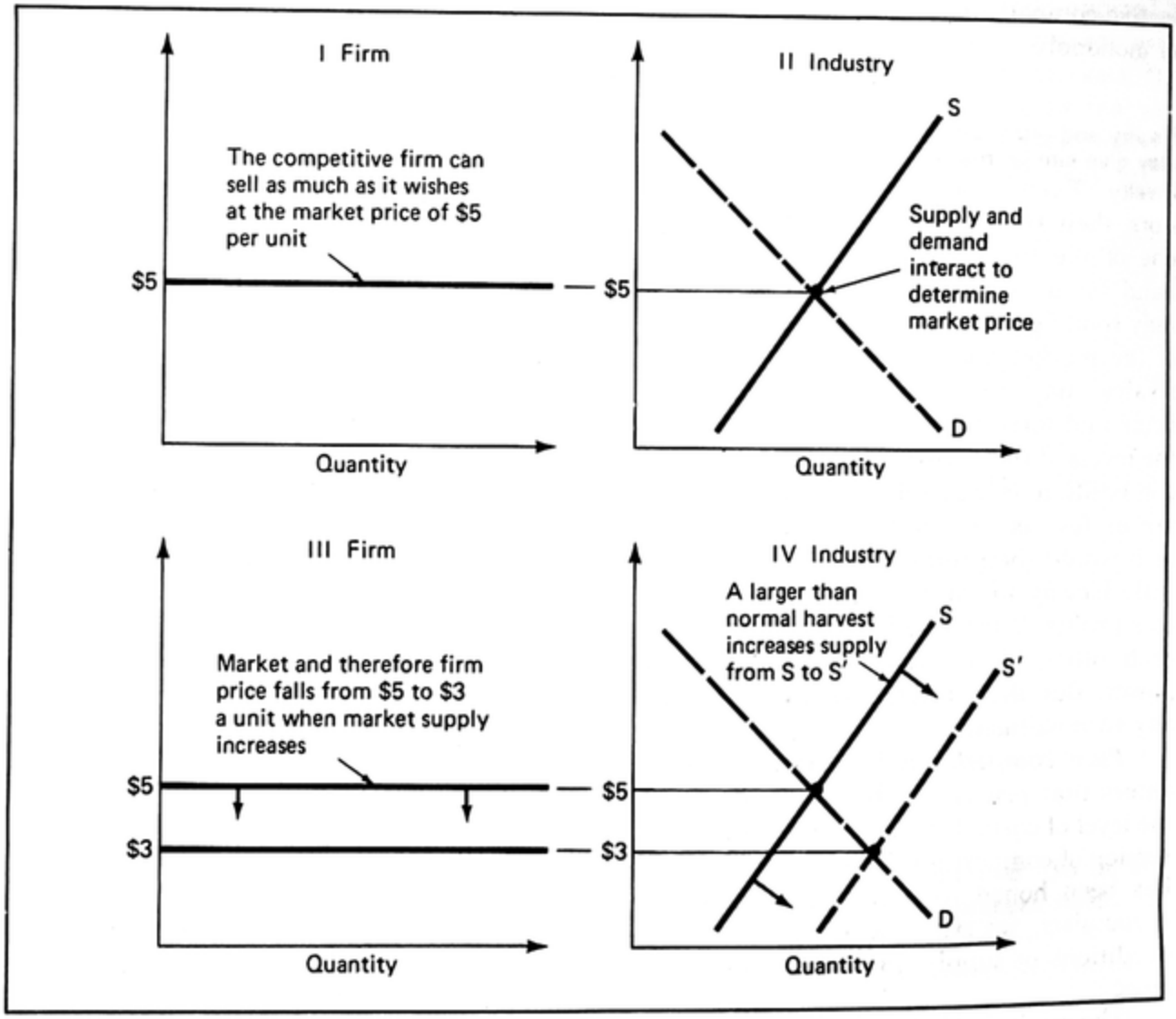


Figure 2 Determination of firm and market price in perfect competition

by an amount equal to market price each time a unit of output is sold. It is represented by a straight line from the origin with a slope equal to $\Delta TR/\Delta Q$, which equals price. With a market price of $5, the firm's MR and TR schedules will be as shown in Figure 3.

Marginal cost

The horizontal demand curve will be superimposed on the firm's cost curves, to locate the output level where profit is greatest. The decisive cost curve is *marginal cost* because it is the precise measure of opportunity cost to the firm.

The strict definition is simple: ***Marginal cost is, at any output level, the added cost of producing one more unit. Therefore, marginal cost is the opportunity cost of the last unit produced.*** It shows how cost varies with that unit of output alone. Marginal cost is, therefore, sharply distinct in logic from average cost.

This sharp variation in short-run marginal costs can be illustrated by two examples. First, consider the marginal cost

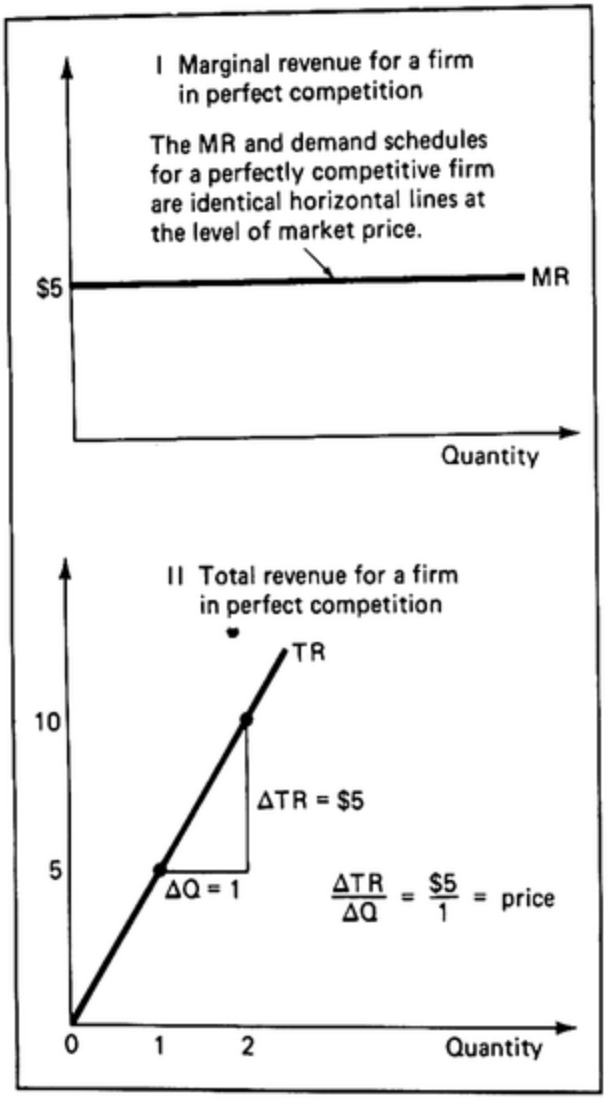


Figure 3 Marginal revenue and total revenue schedules in perfect competition

If market price is $5, the MR schedule is a horizontal line at the level of $5. In perfect competition, then, $P = MR$, as illustrated in Panel I. The TR schedule is a straight line from the origin, with a slope equal to $\Delta TR/\Delta Q$. If $P = \$5$, the slope of the TR schedule will be 5, as is illustrated in Panel II.

of a meal at 3 p.m., when the restaurant is nearly empty. The waiters, cooks, and equipment are already present and paid for. The main extra cost is simply the food itself. So the marginal cost of a meal at slow times might be $1.25. But at 7 p.m., when the restaurant is crowded, an extra meal has much higher cost. An extra waiter and cook may be needed, extra food ordered, extra tables provided, and so forth. These costs might add up to a marginal cost of $15 per meal at peak times.

Second, consider taking a city bus at 11 a.m. or 5:30 p.m. At 11 a.m., the buses are nearly empty; the cost of carrying you is only a little gasoline, perhaps 3 cents' worth. But at 5:30 p.m., the buses are full, and an extra bus may be needed to carry extra riders. That involves an extra driver and other costs.

The firm's short-run supply curve in perfect competition

To derive the ***firm's supply curve,*** simply combine the general rules for profit maximization with the particular characteristics of perfect competition. One profit-maximizing rule is that the firm produce the level of output for which $MR = MC$.

Panel I of Figure 4 shows three MR schedules and one MC schedule for a perfectly competitive firm. If the market price is $5, it will be profit maximizing for the firm to supply 100 units, since that is the point at which MR equals MC. The price–quantity supplied combination of $5 and 100 units would therefore be one point on the firm's supply curve. To generate the entire supply curve, simply vary the price. As the MR curve shifts, other price–quantity supplied combinations can be determined.

For example, if the price is $10, the MR schedule would correspond to MR_2 and the firm would be willing to supply 150 units of output. The price-quantity combination of $10 and 150 units would be another point on the firm supply curve. If the market price were $3, the MR schedule would correspond to MR_3 and quantity supplied would be 70 units. Since these price–quantity supplied points will always be the points at which $MR = MC$, all of the profit-maximizing points must lie along the MC curve. Therefore, the supply curve must be the marginal cost curve. If

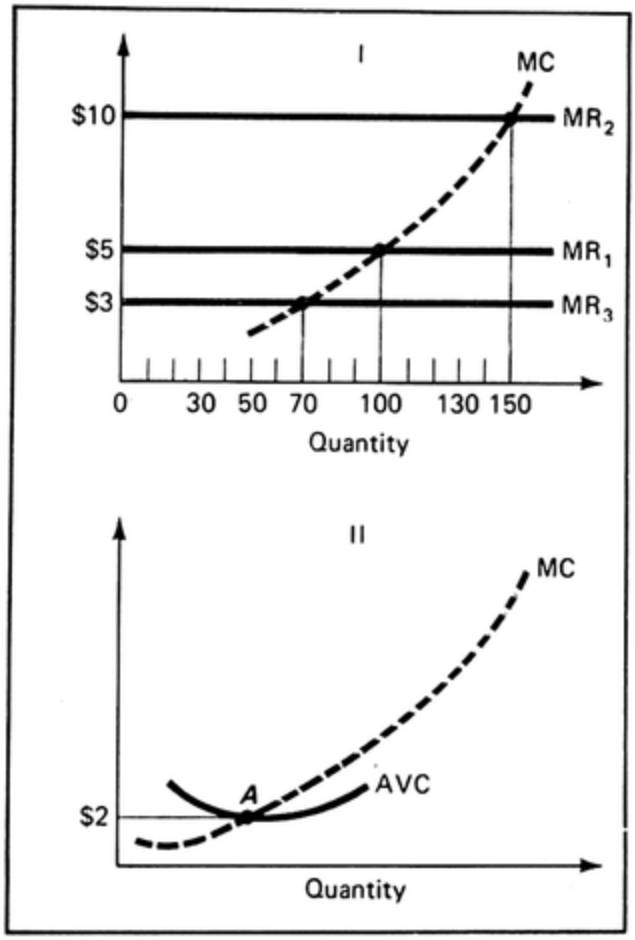


Figure 4 Derivation of a perfect competitor's supply schedule

In Panel I, every quantity supplied–price combination must occur at a *MR = MC* point. Therefore, all supply points must lie along the MC curve. The MC curve, then, is the firm's supply curve over the range in which the firm will operate. In Panel II, only the *A–B* segment of the MC schedule is considered to be the firm's supply curve. The firm will never supply output below $2, which represents minimum AVC.

the firm operates at all, the quantity it wishes to supply will be the quantity indicated on the MC curve at the level of the prevailing market price.

However, the competitive firm's supply curve is not the entire marginal cost curve. Remember that it is profit maximizing for the firm to produce in the short run only if $P \geq AVC$. Any quantity for which the corresponding price is less than AVC will not be supplied. As Panel II of Figure 4 shows, *the perfect competitor's supply curve is the marginal cost schedule at or above AVC*. The dashed portion of the firm's MC schedule in Panel II is not considered part of the firm's supply curve, since the firm will not produce any output at a price lower than its minimum AVC, which in this case is $2.

A more complete analysis

Next we show the profit-maximizing choice in more detail, involving total and average curves as well as marginal ones. Figure 5 presents the conditions for Chapter 7's restaurant *at one time period*, namely 1984. The numerical values for the diagram are in Table 1. The flat (and coincident) demand and marginal revenue curves are derived from the total revenue curve, shown in Panel I. The derivation follows the standard steps, in relating total values to average and marginal values. Average revenue is the slope of the ray through the origin to each point on the total curve. Marginal revenue is the slope of the total curve itself at each point.

Note that the cost and demand conditions are superimposed in both panels of Figure 5. Profit is shown most clearly and simply in Panel I, as the gap between total revenue and total cost. That vertical distance is drawn as a "profit hill." The hill peaks at Point 1, where output is 200,000 meals per year (or 548 meals on the average day). Thus, 200,000 per year is the profit-maximizing level of output. You, the restaurant manager, choose that level, produce it while carefully minimizing the costs, sell the output, and thereby gain profits of $220,000.

The restaurant's choice and its motives are also shown in Panel II. Consider it carefully. Remember that the universal rule is: ***Profits are maximized at the output where marginal cost equals marginal revenue:*** Marginal revenue = Marginal cost. Then the last unit produced is just worth what it cost (in marginal cost) to the firm (in marginal revenue).

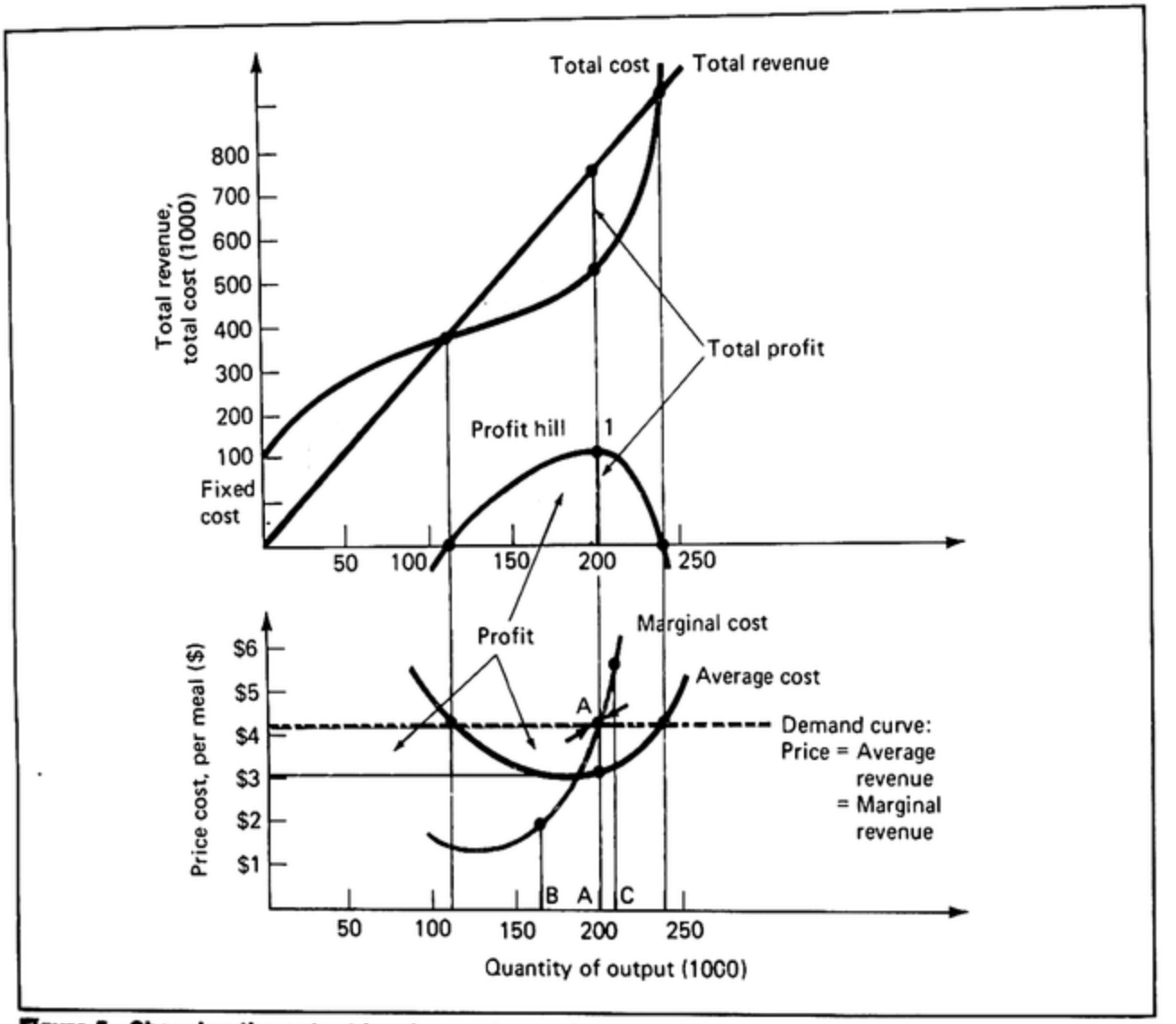


Figure 5 Choosing the output level so as to maximize profits

The total revenue curve has been added in at the top of the figure. Since it is straight, its slope at successive points (marginal revenue) and the slope of the ray from the origin (average revenue, which equals price) are both constant and have the same value. That value is the price, which is shown below as the firm's horizontal demand curve. Such a flat demand curve applies to a perfect competitor, which has to accept the going market price.

Profit is the excess of total revenue over total cost, as shown by the vertical distance in the top panel. The "profit hill" is simply that vertical gap. Below, the average profit per unit is the gap between price (average revenue) and average cost.

Profit is maximized where *price equals marginal cost*. Below that level, marginal output sells for more than it costs. Above that level, the marginal output costs more than its price. Therefore, the firm will move up to, or cut back to, exactly that profit-maximizing level. The resulting profit is the vertical distance in the top panel and the shaded rectangle in the lower panel. It is $1.10 per unit on 200,000 meals, or $220,000.

Under strict competition, marginal revenue is identical to price throughout because the firm's demand curve is horizontal. The demand curve and the marginal revenue curve are the same. Thus, the firm chooses the output level where

Price = Average revenue = Marginal revenue
= Marginal cost.

For competitive firms, therefore, this special condition holds at the profit-maximizing output:

Price = Marginal cost.

Output is set where the rising marginal cost curve cuts the horizontal demand curve.

Table 1 Maximizing profits: The quantities in Figure 5

Quantity of Output (1000 meals per year)	Total Cost (1000)	Total Revenue (Price × Quantity) ($1000)	Total Profit (Total Revenue − Total Cost) ($1000)	Marginal Cost ($ per meal)	Average Cost (Total Cost ÷ Quantity) ($ per meal)	Price and Marginal Revenue ($ per meal)	Average Profit (Price − Average Total Cost) ($ per meal)	Marginal Profit (Price − Marginal Cost) ($ per meal)
0	200	0	−200	—	—	—	—	—
100	465	425	−40	1.60	4.65	4.25	−.40	2.65
120	480	510	30	1.50	4.00	4.25	.25	2.75
140	495	595	100	1.50	3.54	4.25	.71	2.75
160	515	680	165	1.90	3.22	4.25	1.03	2.35
170	535	723	188	2.20	3.15	4.25	1.10	2.05
180	560	765	205	2.60	3.11	4.25	1.14	1.65
190	590	808	218	3.25	3.10	4.25	1.15	1.00
200	630	850	220	4.25	3.15	4.25	1.10	0
210	685	893	208	5.50	3.26	4.25	.99	−1.25
220	755	935	180	7.10	3.43	4.25	.82	−2.85
230	850	978	128	10.00	3.70	4.25	.55	−5.75
240	1020	1020	0	14.50	4.25	4.25	0	−10.25
250	1270	1063	−207	21.00	5.08	4.25	−.83	−15.75

Note how this occurs in the precise conditions of Figure 5. In Panel II, the 200,000 unit at Point *A* is just "worth it." It costs $4.25 to make that meal (that is its marginal cost) and it sells for $4.25, just enough to cover that marginal cost. The 165,000th unit at Point *B* is definitely worth producing, since it costs only $2.00 to make and it sells for $4.25. So all of the units between 165,000 and 200,000 will be produced because they each return a marginal profit. From 200,001 on up, by contrast, each marginal unit causes a marginal loss (shown in the last column of Table 1). Thus, the 210,000th unit, at Point *C*, costs $5.50 but sells for only $4.25: Its marginal loss is $1.25. Producing more than 200,000 units is foolish.

Starting anywhere *away from* the level where price equals marginal cost, the firm can make more profit by changing to that level. That is clear in both panels of Figure 5. Choosing that level is good for the private interest of the firm. It also brings value (the price reflecting consumer preferences) into line with opportunity cost to the economy (marginal cost). We will return to this price = marginal cost condition later in the chapter.

Profit At 200,000 units, the average revenue (price) is $4.25 and average cost is $3.15, as shown in Panel II of Figure 5 and in Table 1. Thus, average profit is $1.10 per unit. Total profit is $1.10 times the number of units (200,000), for a total of $220,000. The profit is shown by the shaded rectangle in Panel II. That rectangle corresponds exactly to the vertical profit distance in Panel I. This profit is "extra" profit. Average cost already includes the cost of capital. Here, the firm is lucky enough to make an extra profit of $220,000 per period in the short run.

Now suppose that the market price for meals drops during the slow summer months to $2.50. In this situation, your restaurant can only do a special short-run version of maximizing profits: minimizing its losses. The restaurant still produces where price equals marginal cost. But now it also applies a special threshold crite-

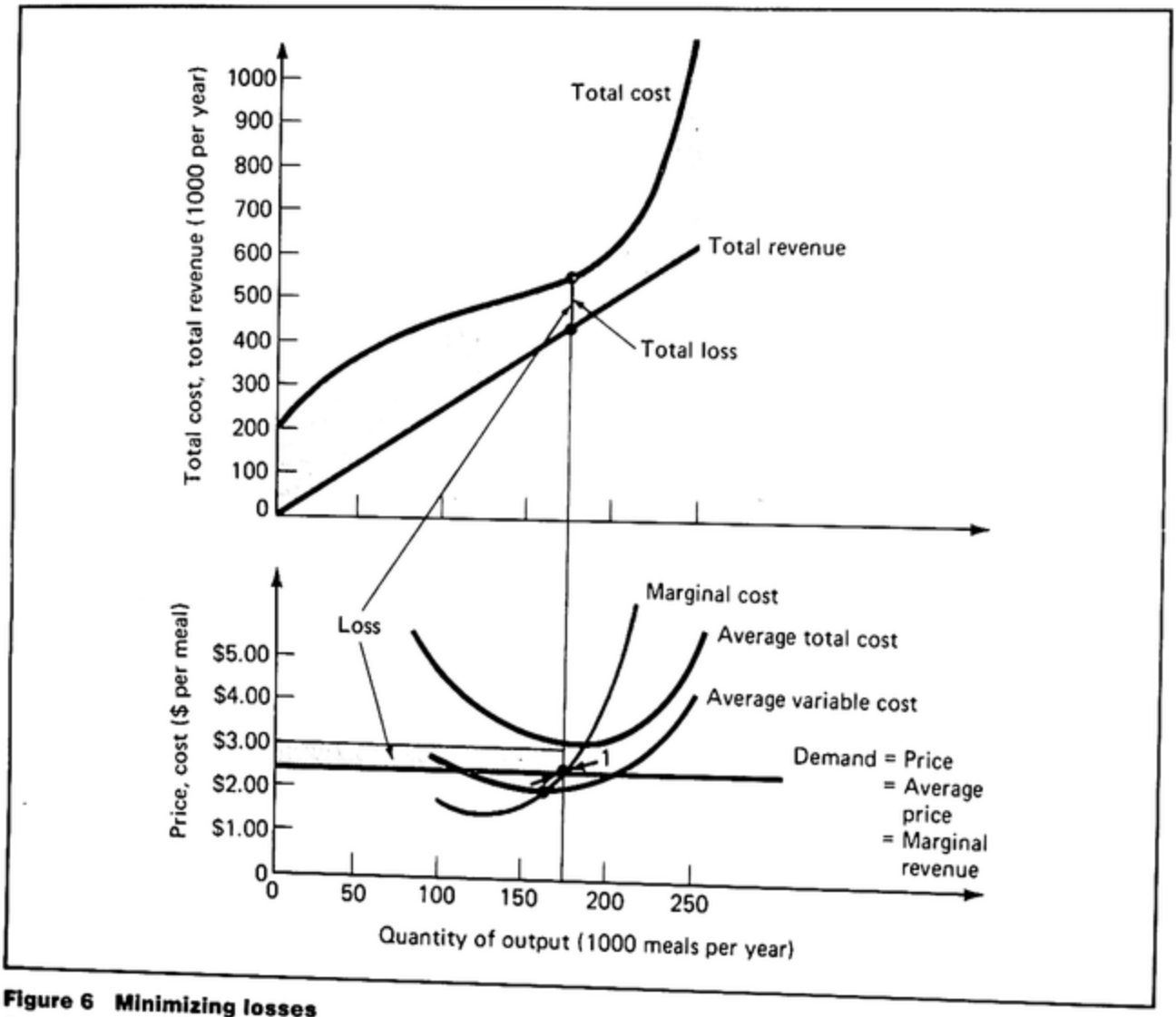


Figure 6 Minimizing losses

Price has now fallen to $2.50 per meal. The total revenue curve has rotated below the total cost curve at every point. No profit is now possible, but at least the losses can be minimized.

In the lower panel, price equals marginal cost at 175,000 meals per year. That is the best output level. At lower output levels, the firm can make marginal profits by expanding. At outputs above 175,000, the restaurant makes marginal losses and should cut back.

The vertical loss *distance* in the upper panel corresponds precisely to the shaded loss *rectangle* below.

rion: ***Price must cover average variable cost.*** If the price equals or exceeds average variable cost, then the firm keeps producing in the short run, until its fixed costs expire. Then all of the costs become variable, and price must cover average total cost. If at any time price is below average variable cost, the firm closes down at once.

Therefore, the competitive firm's supply curve is the part of the marginal cost curve *that is at or above the average variable cost curve.*

This fits both common sense and the familiar business wisdom: (1) to keep producing in the short run it the price covers at least your current (or "out-of-pocket") expenses, but (2) to shut down in the long run unless price covers all average costs. It is illustrated in more detail in Figure 6. Average variable cost goes as low as $1.97 per meal. At a $2.50 price, you will definitely stay open during the summer.

How many meals will you supply? The answer is shown by marginal cost, just as before. Where it equals price, losses are minimized. That occurs at a rate of 175,000 meals per year, as shown by Point 1 in Panel II of Figure 6. Think of it as an av-

Table 2 Minimizing losses: The quantities in Figure 6

Quantity of Output (1000 meals per year)	Total Cost ($1000)	Variable Cost (Total Cost − Fixed Cost) ($1000)	Total Revenue (Price × Quantity) ($1000)	Total Profit (Total Revenue − Total Cost) ($1000)	Marginal Cost ($ per meal)	Average Total Cost (Total Cost ÷ Output) ($ per meal)	Average Variable Cost (Variable Cost ÷ Output) ($ per meal)	Price and Marginal Revenue ($ per meal)	Average Profit (Price − Average Total Cost) ($ per meal)	Marginal Profit (Price − Marginal Cost) ($ per meal)
0	200	0								
100	465	265	250	−215	1.60	4.65	2.65	2.50	−2.15	.90
120	480	280	300	−180	1.50	4.00	2.33	2.50	−1.50	1.00
140	495	295	350	−145	1.50	3.54	2.11	2.50	−1.04	1.00
160	515	315	400	−115	1.90	3.22	1.97	2.50	−.72	.60
170	535	335	425	−110	2.20	3.15	1.97	2.50	−.65	.30
180	560	360	450	−110	2.60	3.11	2.00	2.50	−.61	−.10
190	590	390	475	−115	3.25	3.10	2.05	2.50	−.60	−.75
200	630	430	500	−130	4.25	3.15	2.15	2.50	−.65	−1.75
210	685	485	525	−160	5.50	3.26	2.31	2.50	−.76	−3.00
220	755	555	550	−205	7.10	3.43	2.52	2.50	−.93	−4.60
230	850	650	575	−275	10.00	3.70	2.82	2.50	−1.20	−7.50
240	1020	820	600	−420	14.50	4.25	3.42	2.50	−1.75	−12.00
250	1270	1070	625	−645	21.00	5.08	4.28	2.50	−2.58	−18.50

erage of 480 meals per day during the low-demand summer season. Table 2 corroborates the answer. Note the column for marginal profit. Added units make a profit up to about 175,000 meals per year. You move up to 175,000 and stay open. Added meals cause extra losses. The Total Profit column clinches it. Losses are the least possible between 170,000 and 180,000 meals per year. These losses are shown in Panel I of Figure 6 by the vertical distance between the total revenue and total cost curves. That distance is shortest at 175,000 meals per year.

The basic rule is that revenues must exceed corresponding costs. *In the short run, average variable costs are the test. In the long run, all costs are variable, so that average total costs become the minimum test.* Generally, whatever the time period, output is set where marginal revenue equals cost.

Figure 7 shows precisely how the firm's short-run supply curve is the part of its marginal cost curve that lies above average variable cost. At prices below variable cost, the firm produces nothing. At higher prices, output is set along the marginal cost curve.

The short-run market supply curve is a summation

The market includes all active buyers and sellers of the good. Within that market are all the firms that might produce and sell at the going price. Our typical firm is one.

Each firm's marginal cost curve is its supply curve. Added up horizontally, these curves form the ***supply curve for the whole market.*** **The short-run market supply curve is the summation of the firms' short-run marginal cost curves.** Figure 8 shows the horizontal summation. At low prices, some firms supply nothing, while others supply moderate amounts. As prices rise, some firms begin producing—as price rises above their average variable cost—while the firms already producing raise their levels of production. Since marginal cost curves slope up (reflecting diminishing marginal returns), so does the market

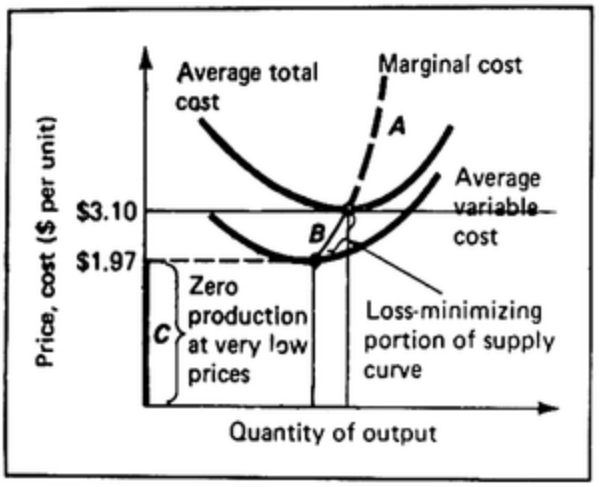


Figure 7 The short-run supply curve of a competitive firm

The supply curve includes Portions *A*, *B*, and *C* in the short run. Price can fall to $1.97 before the firm shuts down immediately. In the *B* range, the firm is taking temporary losses.

supply curve. In short, the shape and position of the supply curve reflect the technology of the industry, as it is embodied in the marginal and average variable costs curves.

Shifts in the firm and industry supply curves

The firm supply curve must shift if its MC curve shifts. Shifts in the MC curve will occur if the variable costs of production change because of changes in either input prices or technology. If, for example, input prices fall, a firm's MC curve will shift down to the right. The firm will be willing to supply more output at every price (or to supply a given level of output at a lower price) because it is profit maximizing to do so.

Various shifts in the firms' MC curves are explored in Panels I and II of Figure 9. The industry supply curve will shift if the firms' MC curves are shifting or if the number of firms in the industry changes. If the number of firms in the industry increases, then more marginal cost curves must be added to arrive at the industry supply curve. The industry supply curve would shift right, showing that more output will be offered at every price. Shifts in the industry supply curve are also explored in Panels III and IV of Figure 9.

Three points about the supply curve should be carefully noted. First, the firm and industry supply curves presented here apply only to perfectly competitive firms and industries because the firms supply

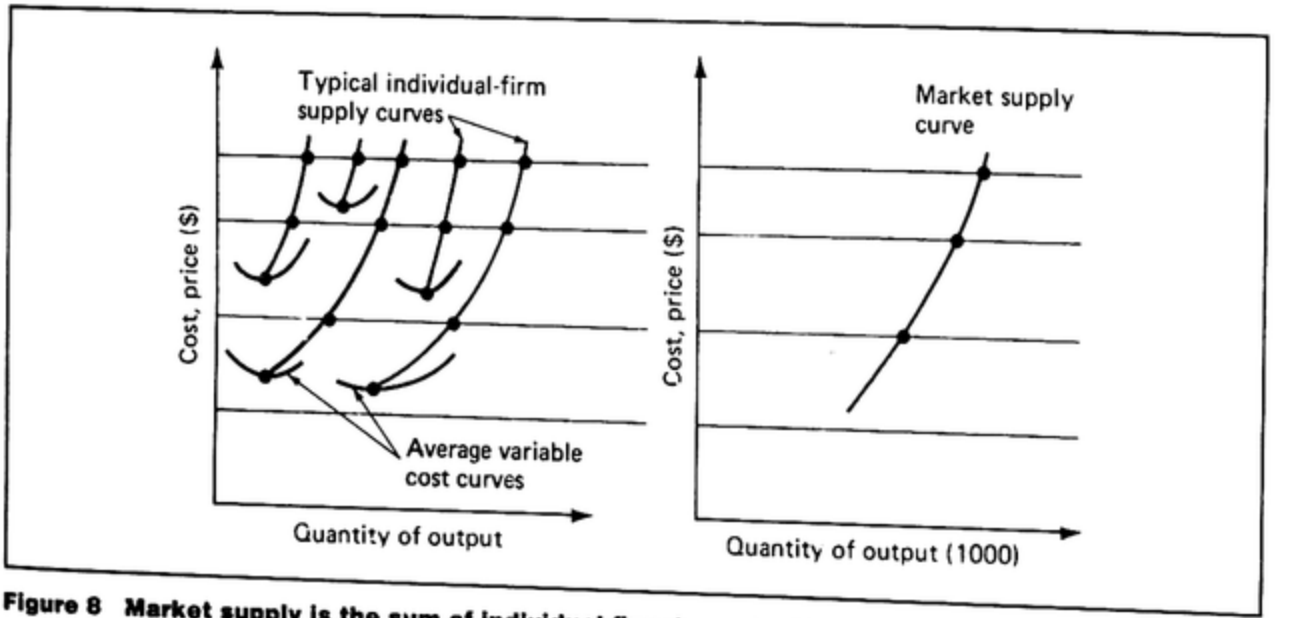


Figure 8 Market supply is the sum of individual firms' supply: Short run

At each price, the firms will act along their own supply curves. Their outputs make up the market's total supply. The market supply curve reflects the shape of the firm's various marginal cost curves. Since they slope up, the market supply curve slopes up.

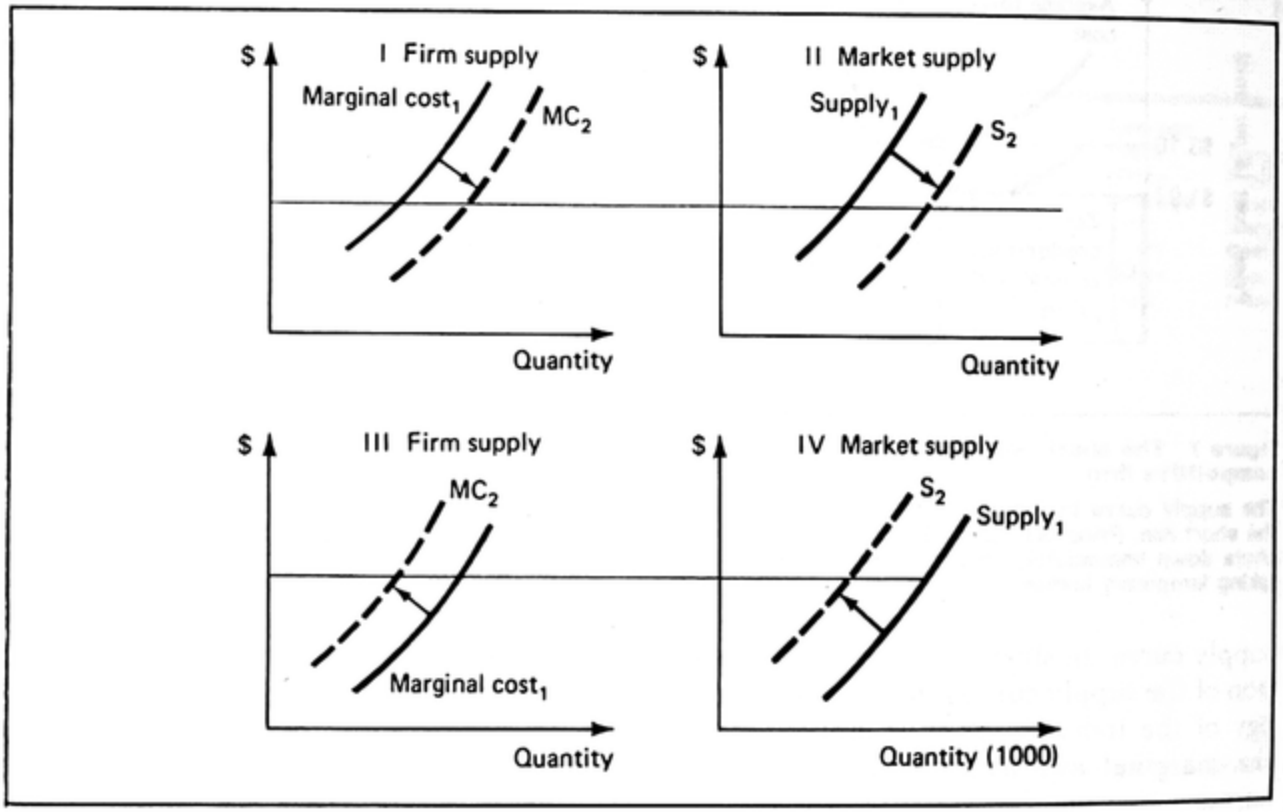


Figure 9 Shifts in firm and market supply curves

In Panels I and II, there has been a decrease in the price of inputs, or a technological improvement. Both the firm's supply surve and the market supply curve shift down.

In Panels III and IV, there has been a rise in input prices, and a decrease in the number of firms. Both the firm's supply curves and the market supply curve shift up.

curve was derived on the assumption of a horizontal MR curve. Since the horizontal MR curve is a result of the specific assumptions of perfect competition, namely that an individual firm is a price taker, the resulting supply curve pertains only to perfect competition.

Second, not all cost changes will affect the firm and industry supply curve. Only cost changes that affect variable and therefore marginal costs can cause the supply curves to shift. If fixed costs increase, the firm's MC or supply curve will not shift. The industry supply curve may shift eventually if the change in fixed costs results in a change in the number of firms.

Third, the firm and industry supply curves discussed here are *short-run* supply curves. Since the firm's MC curves are associated with a particular-size plant, fixed factors are involved. The industry supply curve is also based on fixed factors, since it represents a fixed number of firms. If the number of firms changes, the supply curve will shift.

Short-run and long-run equilibrium in perfect competition

The short-run and long-run equilibrium conditions for any type of firm can be derived by combining the characteristics of the firm's market with the rules for profit maximization. The rules for profit maximization will, of course, be the same regardless of market structure. It is the different characteristics of each market or industry that will lead to very different equilibrium outcomes.

In perfect competition, the key characteristics to remember are that the firms in the industry are price takers, and that there is freedom of entry and exit. The profit-maximizing rules are that the firms must cover variable costs in the short run and total costs in the long run. Firms must also produce at the point where $MR = MC$.

The three panels in Figure 10 show three possible profit-loss situations for a perfectly competitive firm. Panel I shows a firm taking a loss, with $P < ATC$. Panel II shows a firm making a normal return, with $P = ATC$. Panel III shows a firm making a profit, with $P > ATC$. Which of these situations is compatible with *short-run* equilibrium?

Equilibrium, remember, means a state of rest or balance, a state of stability. For a firm to be in equilibrium, it must be profit maximizing. After all, if the firm is not doing as well as it can, it will certainly have the incentive to make changes. Therefore, for ***short-run equilibrium,*** the short-run profit-maximizing rules of $P \geq AVC$ and $MR = MC$ must be satisfied. In each of the three situations depicted by the panels in Figure 10, these short-run profit rules are satisfied. Therefore, all three situations of profit, loss, and normal return are compatible with short-run equilibrium. In none of the situations can the firm improve its position. The loss shown in Panel I would not make the firm particularly happy. If $P < AVC$, the firm would immediately shut down. However, as long as $P \geq AVC$, it is best for the firm to continue to operate: It will at least minimize its losses by paying off part of its fixed cost.

In the long run, however, the situation is different. The profit-maximizing rules are now that $P \geq ATC$ and $MR = MC$. Since $P < ATC$ in Panel I, that cannot represent a ***long-run equilibrium.***

In Panel III, $P > ATC$. Could this situation represent a long-run equilibrium? To answer that question, you need to think about the specific characteristics of perfect competition. In perfect competition, there is freedom of entry and exit. If the firms in perfect competition are making a profit, with $P > ATC$, other firms will enter the industry. As that happens, the industry supply curve will shift right and market price will begin to fall. Each firm will find its own MR curve or demand curve shifting down in response to the decreased market price, as Figure 11 shows.

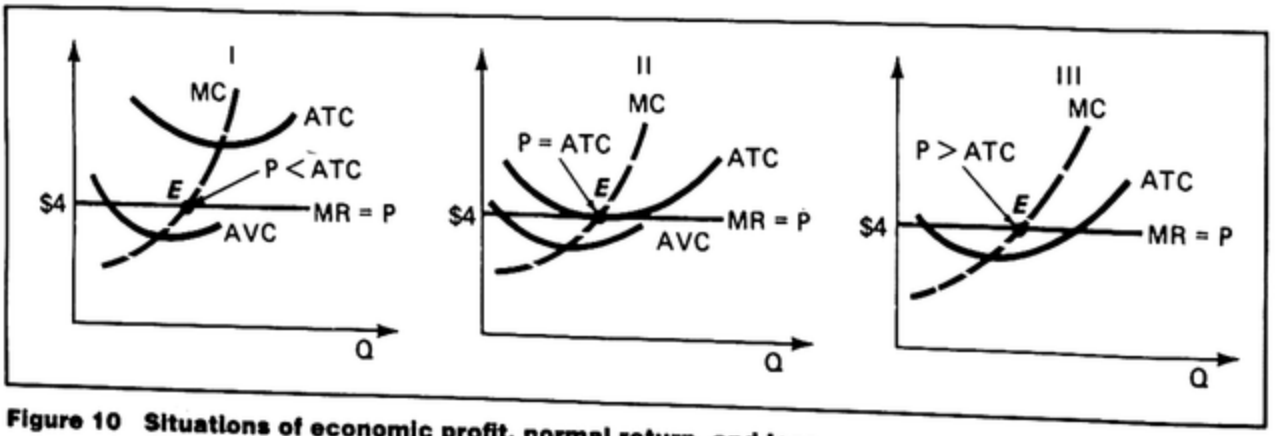


Figure 10 Situations of economic profit, normal return, and loss

At Point *E* in Panel I, $MR = MC$ and $P > AVC$. The firm is in short-run equilibrium. Panel I does not represent a long-run equilibrium, however. In the long run, losses will cause exit. At Point *E* in Panel II, $P = ATC$ and $MR = MC$. The firm is in short-run equilibrium. Panel II also represents long-run equilibrium. With $P = ATC$, there is no incentive for entry or exit. At Point *E* in Panel III, $MR = MC$ and $P > ATC$. The firm is in short-run equilibrium. Panel III does not represent long-run equilibrium, however. With $P > ATC$, entry will occur.

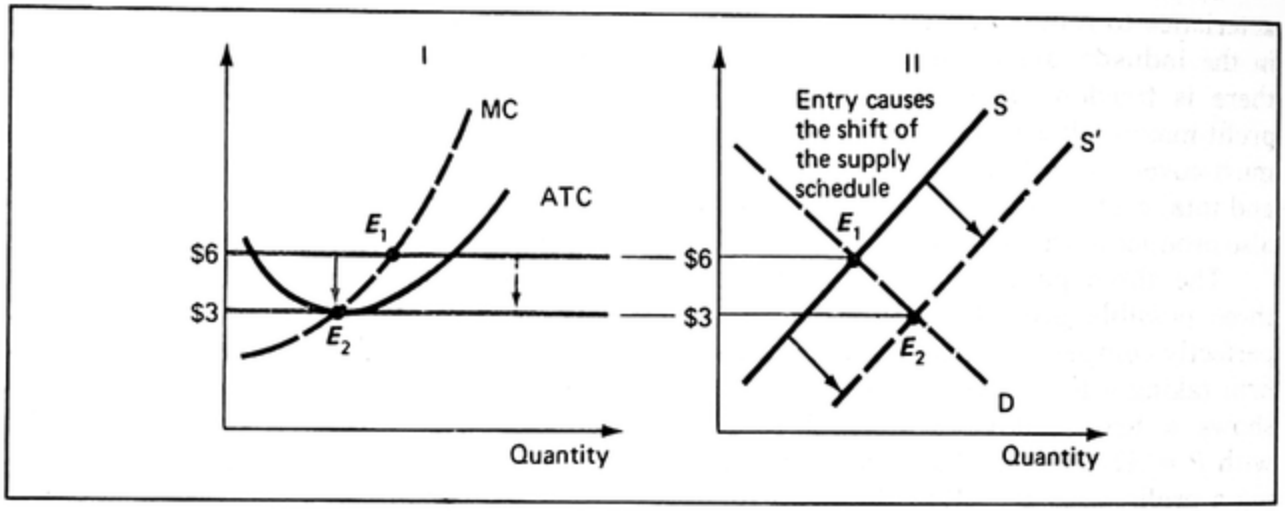


Figure 11 Adjustment toward a long-run equilibrium

At E_1 in Panel I, the firm is making a profit, with price of $6 greater than ATC. Profit attracts new firms, and their entry causes the industry supply schedule to shift to the right, as in Panel II. As the supply schedule shifts right, the market price will fall, shifting the firm's MR or demand curve downward. Entry will cease only when market price equals minimum ATC. This occurs at a market price of $3. At E_2, $P = ATC$ and $MR = MC$. Each firm is maximizing profits and there is no incentive for entry or exit. E_2, therefore, represents a long-run equilibrium.

These shifts will continue as long as $P > ATC$. Since changes are not compatible with an equilibrium, it is clear that economic profits cannot exist in long-run equilibrium for a competitive industry. Only with $P = ATC$ is there no incentive for firms to enter into or exit from the industry. Neither profits nor losses are being made then. Thus, for perfect competition, the two long-run equilibrium conditions are: (1) $P = ATC$, and (2) $MR = MC$.

Because the assumptions about competition are so specific, we can be very precise about the long-run equilibrium outcome in perfect competition. Not only must firms be making a normal return with $P = ATC$, *but ATC must be at its minimum point.* This result stems from the price-taking assumption of perfect competition. Examine Figure 12, which shows a U-shaped long-run ATC curve. Suppose that a competitive firm were operating at Point *A*. Could this possibly be a stable or equilibrium position? No, it could not. Since price per unit is unchanged by the

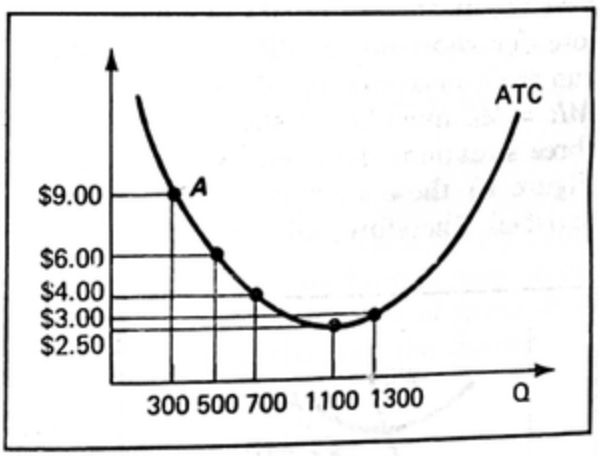


Figure 12 Long-run average total cost

amount the firm produces, the firm has a clear incentive to increase plant size until it is producing at the lowest or minimum ATC. Therefore, in the long run, when the firm can adjust all factors, it will clearly have an incentive to adjust its technology and its output levels until it is producing at minimum ATC. Figure 13 portrays a perfectly competitive firm and industry in long-run equilibrium. Notice how all of

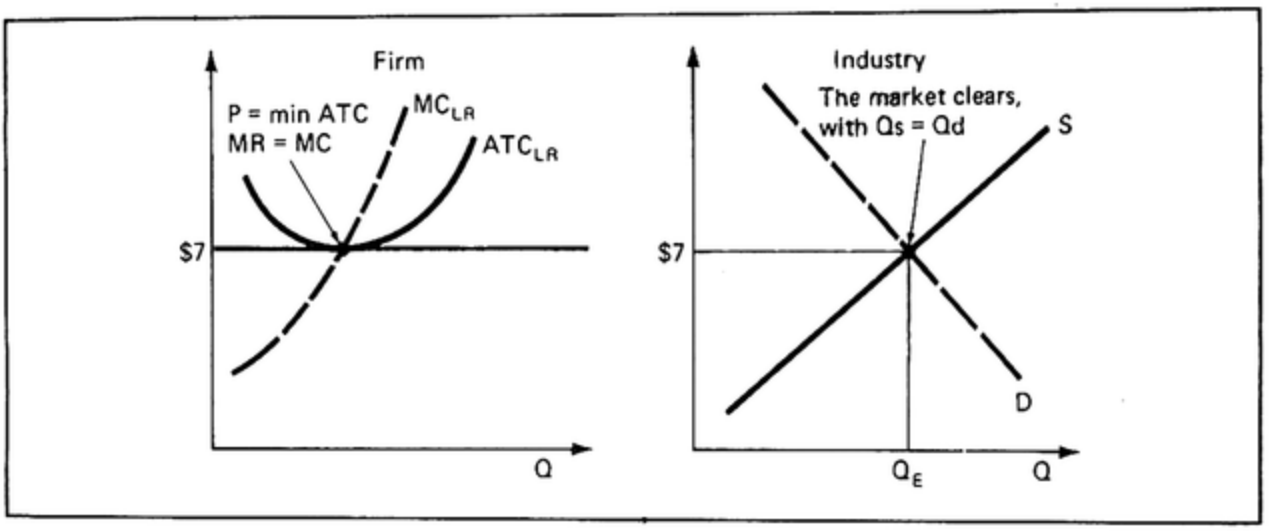


Figure 13 Long-run equilibrium for a perfectly competitive firm and industry

the equilibrium conditions are illustrated in one diagram. At the level of output produced, $MR = MC$, and P = minimum ATC.

Long-run supply

The firm's long-run supply curve

The firm's ***long-run supply curve*** is its *long-run marginal cost curve* lying above average total cost. That is shown in Figure 14. These cost curves were derived in Chapter 8. Recall that the long-run curve is flatter than the short-run supply curve, reflecting the wider choices available in the long run, when no input is fixed. The same rule applies in setting output. At prices below average variable cost (which is now *identical to average total cost* because all costs are variable in the long run), output is zero. The firm is closed down. Above that price, the marginal cost curve is the supply curve. In short, supply is precisely determined by price and cost.

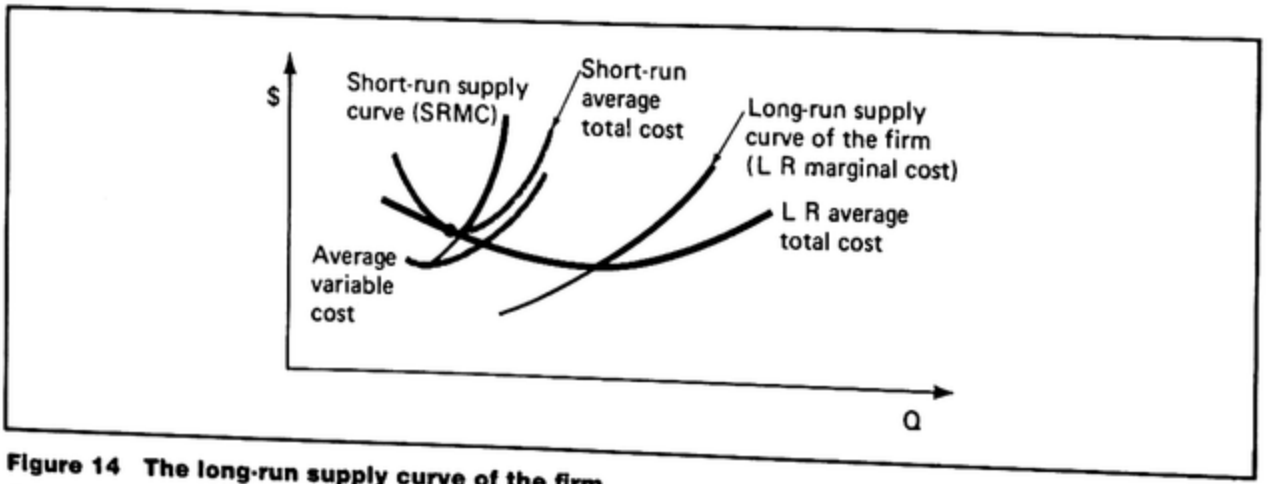


Figure 14 The long-run supply curve of the firm

For a given size of plant, the short-run cost curves show the firm's choices. But in the long run, the firm's supply curve is the long-run marginal cost curve, as shown.

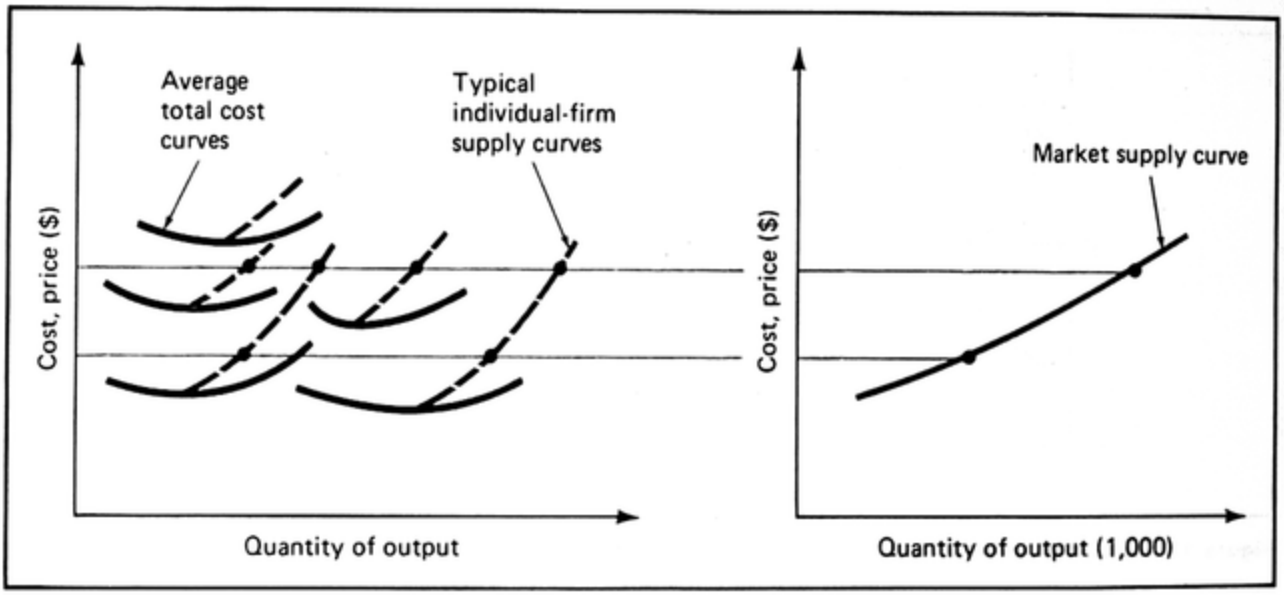


Figure 15 Market supply in the long run

Now the firms' long-run supply curves are more elastic than their short-run supply curves, because adjustments can be made over a larger range of inputs in the long run.

The long-run market supply curve

As in the short run, the long-run supply curve for the market is the horizontal summation of the firms' supply curves. The summation is shown in Figure 15. The long-run supply curve is more elastic, because the firms' long-run marginal cost curves are more elastic. Also, the long run permits the entry of new firms, which augment the supply as price rises. Nonetheless, both curves are derived by the same basic process of horizontal addition.

Efficient allocation under competition

Now consider what the analysis in this chapter has established. The steps have been relatively simple, primarily a horizontal adding up of marginal cost. Yet, the result is profound. In each market, supply occurs as the result of consistent patterns of cost and price. The quantities supplied are precisely governed by the going market prices and the internal costs of all firms in the market. Higher prices induce greater levels of output; lower prices cause production to be cut back. These outcomes are intuitively likely. Here they have been explained by exact, consistent logic.

Moreover, the outcome is in line with *efficient production*. Efficiency is used in economics to define a set of conditions. Several of those efficient conditions have already been defined along the way, in explaining the competitive result. We will now list each of them:

Avoiding unnecessary cost Each firm attains the costs shown by the average cost curve, rather than incurring higher costs. Whatever output level is chosen, average cost is as low as possible. This is termed ***x-efficiency***, as noted in Chapter 8.

Optimum scale Each firm also chooses the output level at which average total cost is at its minimum, in the long run. This level is termed optimal scale, for it is the size at which average cost levels are as low as possible.

Allocative efficiency ***Allocative efficiency*** deals with the pattern of inputs and outputs among all firms and markets. Even if production efficiency and optimal scale are reached in every firm, there might still be too much of some outputs produced and too little of others. The allocation of resources among goods would be inefficient. What is the right pattern of goods? The general answer is: ***that allocation in which price equals marginal cost for every firm.*** Recall that price represents the value of a good to consumers, shown by what they will pay for it at the margin. Marginal cost is the opportunity cost of the good to the firm. When price equals MC, the firm's own choice is efficient, for the last unit is just worth its cost.

For society as a whole, too, price equals MC defines the best level for each output. Price is the *social* value of the good at the margin: What people will pay for the good is a measure of its value to the general population. Private and social value are identical. As for cost, marginal cost measures the *social* opportunity cost of the last unit of the good. It is the degree of additional sacrifice (in effort, resources, etc.) that the workers must make to sustain the current level of output. When price equals marginal cost, then the value of the added output just equals the sacrifice to produce it.

To clarify efficient allocation, consider a departure from it. Suppose all firms meet the $P = MC$ condition except one, where output is "too small" and price exceeds marginal cost. By expanding production, this firm will create more value (price) than cost (marginal cost). It will

Table 3 Basic conditions of efficient equilibrium

Equilibrium Condition	Cause
Short Run	
Price > Average variable cost	A firm will only operate in the short run if it can cover the costs of staying open (the variable costs pay off part of its fixed costs).
Marginal revenue = Revenue cost	Condition necessary for the profit-maximizing level of output.
Price = Marginal cost	Price equals marginal cost for the competitive firm.
Long Run	
Price = Average total cost	*Freedom of entry and exit:* Since firms are free to enter or leave the industry, equilibrium can only occur if there is no incentive for firms to enter or exit. This implies that a normal return (economic profits = 0) is being made by the firms in the industry, since losses would cause exit and profits would cause entry.
Price = Marginal cost	*The condition of profit maximization under competition:* It also equals average cost (the condition of normal return). So price must equal average cost at the point where average and marginal costs are equal, at minimum average total cost.
Marginal revenue = Marginal cost	Condition necessary to achieve the profit-maximizing level of output.
Price = Marginal cost	Price equals marginal cost for the competitive firm.

take resources from uses where they brought no net gain and, in this firm, create a net gain. The total value of production will rise, even if only slightly. When adjustments no longer add net value, then allocation is efficient.

Thus, condition of efficiency will be enforced by competition. Each firm maximizes profits by applying precisely the same $P = MC$ condition that defines efficient allocation for the entire economy. Table 3 (on page 195) summarizes the conditions in detail. If they are reached by all firms, then allocation is efficient throughout the economy.

Limits Neoclassical economists have known the efficiency benefits of a competitive economy for nearly a century. But the process also has various limits, which can cause the outcomes to deviate from social goals:

1. *Distribution.* The distribution of income and wealth may be unfair, even if allocation is efficient.
2. *Technological progress.* The rate at which technology is improved is also largely outside the pure competitive process.
3. *Market prices and costs may deviate from the true social values.* Such deviations would mean that the competitive outcome would not be socially efficient after all. Common examples are the air and water pollution created by some "efficient" factories.
4. *Monopoly.* Competition may be replaced by monopoly in one or more markets. That distorts allocation away from efficiency.

The benefits and limits of the competitive outcome are presented in detail in the chapter on General Equilibrium.

Summary

This chapter first establishes the general rules that all profit-maximizing firms must follow. It then applies these rules to the model of industry behavior known as perfect competition. Short-run and long-run equilibrium conditions for the industry are then derived.

1. A firm should produce in the short run if it is at least covering its operating costs, with $P \geq AVC$. In the long run, a firm will only produce if it can cover its economic costs, with $P \geq ATC$.
2. A profit-maximizing firm should always produce the level of output for which $MR = MC$. If it is producing at a point where $MR > MC$, it should increase output, since each additional unit of output will add more to revenue than to cost, thereby increasing the firm's profit or reducing its losses. If the firm is producing at a point where $MC > MR$, it should decrease its level of output, since each additional unit is adding more to cost than to revenue, thereby reducing the firm's profit or increasing its losses.
3. Competition may be pure or imperfect. But even a two-firm rivalry may approach the results of pure competition, as long as collusion is prevented.
4. The two key assumptions of the model of market structure known as perfect competition are: (1) the firms in the industry are price takers, accepting the market price as given; (2) there is freedom of entry and exit.
5. In perfect competition, an individual firm's demand and marginal revenue curves are both represented by a horizontal line at the level of market price.

6. The short-run supply curve of a competitive firm is its marginal cost curve at or above AVC. The short-run supply curve for the industry is derived by summing horizontally the firm supply schedule. If the firms' MC curves shift or if the number of firms changes, the industry supply schedule will shift.
7. A competitive firm in *short-run* equilibrium will produce with $P \geq AVC$ and $MR = MC$. It may be making a profit, loss, or normal return.
8. A competitive firm in *long-run* equilibrium will be producing at a point where P = minimum ATC and $MR = MC$.
9. The market supply curve is the horizontal summation of the firms' supply curves.
10. The results of changes that lead to disequilibrium can be determined by combining profit-maximizing rules with the characteristics of a specific industry.
11. A long-run supply schedule that allows for changes in the number of firms can also be derived. The slope of the long-run supply curve will depend upon the behavior of input prices as entry and exit of firms occur.

Key concepts

Rules of profit maximization
Marginal revenue
Competition
    rivalry
    pure competition
    perfect competition
Firm demand curve
Firm supply curve
Supply curve for whole market
Firm and industry equilibrium
    short-run
    long-run
Long-run supply curve
Production efficiency
Allocative efficiency

Questions for review

1. Given the information in the following situations, can you determine if the statements are true or false? Explain your answer.
    a. A firm is selling 5,000 units at $5 each. Its total costs equal $20,000. Therefore, the firm must be profit maximizing.
    b. The last unit that the firm produced added $5 to revenue and $3 to cost. The firm should continue to operate.
    c. A firm is selling 1,000 units of output at $3 a unit. Its average total costs are $2,500. The firm should shut down.
    d. The last unit a firm produced added $3 to revenue and $6 to cost. The firm should operate in the short run but not in the long run.
    e. A firm is producing at the point at which $MR = MC$. The firm must be making economic profits of zero.

2. Consider the following list of firms. Which firms would not satisfy the conditions of perfect competition. Explain.

GM	NBC
wheat farmer	A&P
Bethlehem Steel	Harvard University
bookstore in your town	local bank

3. Consider the following statements:
   a. A firm in perfect competition can sell as much output as it wishes at the prevailing market price.
   b. There was an exceptionally abundant wheat harvest that year. Every farmer took larger quantities of wheat to market and therefore received a lower price.

   Are these two statements contradictory? Why or why not? Explain.
4. Each of the following statements is incorrect because there is vital information missing. What must be added to each statement to make it correct?
   a. The demand curve in perfect competition is downward sloping.
   b. The supply curve of a perfectly competitive firm is its marginal cost schedule.
   c. If the number of firms in a competitive industry changes, the industry supply schedule will shift to the right.
5. State which characteristics of a perfectly competitive industry will cause each of the following long-run equilibrium conditions. Explain your answers.
   a. $P = ATC$
   b. $MR = MC$
   c. production at minimum ATC

10

Monopoly

As you read and study this chapter you will learn:

- the varieties of market forms in which monopoly power occurs
- the nature of monopoly and its effects on efficiency, equity, and other social values
- the possible role of scale economies as a cause of monopoly power
- several case studies of monopoly

Whenever friends settle down to play Monopoly, their little board game is much like the endless pursuit of monopoly in real industrial markets. Each player tries to amass as much of the property—real estate, railroads, and utilities—as possible. The player who gets all of one set of properties can force the others to pay much higher rents on them.

The players all strive to monopolize the properties, to extract high profits from them, and, in the process, to bankrupt one another. Eventually one player, the successful monopolist, emerges victorious and the rest are impoverished.

In real industrial life, too, firms strive to gain monopoly power in their markets and to reap its rewards. Pure monopoly—from the Greek word *monopolion*, meaning exclusive sale—has a simple meaning: the control of all sales in a market. Even though a firm has less than 100 percent of the market, it may still have an important degree of monopoly power.

If all firms try to gain a monopoly and yet none of them manages to prevail, then the result is a continuous, healthy competitive process. The aspiring monopolists neutralize one another. When the resulting competition is vigorous, the benefits will be large.

In contrast with competition, monopoly usually causes economic harms. It may distort the allocation of resources, cause waste and inefficiency, and shift wealth unfairly. From the earliest times, rulers have issued countless laws to deal with the harms and distortions caused by monopoly power. Such a serious problem needs careful study. Therefore, this chapter and the two that follow it present the causes and effects of, and the cures for, monopoly in some detail.

This chapter begins by showing the main types of monopoly and competition. Next, we show the distinctive forms that monopoly power takes, and then the main ways by which monopoly can be created. Then, we present the effects of monopoly, contrasting them with the competitive outcomes shown in Chapter 9. Finally, we discuss some case studies of prominent monopolies. They include Standard Oil from 1880 to 1910 and electrical generation in the early twentieth century.

Varieties of monopoly and competition

Pure monopoly is a powerful device for gaining wealth, as we shall soon see. Pure monopolies, however, are highly unusual in modern industry. The polar opposite of pure monopoly is pure competition, where no firm has any control over price. It, too, is unusual.

Most markets lie between these extremes, with some degree—slight, modest, or large—of ***monopoly power*** (also called ***market power***). There are many subtle gradations and varieties of monopoly power, but debates in the literature have settled on several main categories. Table 1 lists these classes, with their main features. The table also suggests which parts of the economy each category is commonly found in.

All real markets fit somewhere into Table 1. The categories shade into one another, rather than being sharply separate, so that some industries are on the fence. For example, the automobile industry has features of both a tight oligopoly (with three leading firms) and a dominant-firm case (with General Motors holding about 45 percent of the market). Moreover, each of the categories in Table 1 covers a range of conditions, rather than just one form.

Whether monopoly is complete or partial, the same basic analysis applies. Monopoly produces certain effects, whose strength depends on the degree of monopoly power. When monopoly power is strong, the effects are strong; where it is weak, monopoly power gives only mild effects. Both in this chapter and in the next, therefore, keep in mind that the analysis is relevant to all degrees of monopoly power.

Monopoly and its effects

The first task is to learn to recognize monopoly by its usual forms.

The characteristics of monopoly

Monopoly can exist when the firm's demand curve slopes down, rather than being horizontal as it is for the purely competitive firm. The down-slope gives the firm a range of choice in setting the price for its product. Figure 1 shows both a down-sloping demand curve *(A)* and a horizontal one *(B)*. The firm with Demand Curve *A* can choose Point 1, with its output and price levels. It could also set a

Table 1 ***Types of markets, shading over from pure monopoly to pure competition***

Market Type	Main Condition	Familiar Instances
Pure monopoly	One firm has 100 percent of the market	Electric, telephone, water, bus, and other utilities; patented drugs
Dominant firm	One firm has 40–100 percent of the market and no close rival	Soup (Campbell), razor blades (Gillette), newspapers (most local markets), film (Eastman Kodak), hospitals
Tight oligopoly	The leading four firms, combined, have 60–100 percent of the market; collusion among them to fix prices is relatively easy	Copper, aluminum, local banking, TV broadcasting, light bulbs, soaps, textbook stores
Loose oligopoly	The leading four firms, combined, have 40 percent or less of the market; collusion among them to fix prices is virtually impossible	Lumber, furniture, small machinery, hardware, magazines
*Monopolistic competition**	Many effective competitors, none with more than 10 percent of the market	Retailing, clothing
Pure competition	Over 50 competitors, all with negligible market shares	Wheat, corn, cattle, hogs, poultry

Source: W. G. Shepherd, *The Economics of Industrial Organization* (Englewood Cliffs, N.J.: Prentice-Hall, 1979), Chaps. 4, 9, 10. Adapted by permission of Prentice-Hall, Inc., Englewood Cliffs, N.J.

*The phrase (coined by E. H. Chamberlin in 1932) means virtually complete competition, but with a moderate degree of differences among products. See the second main section of the next chapter for discussion.

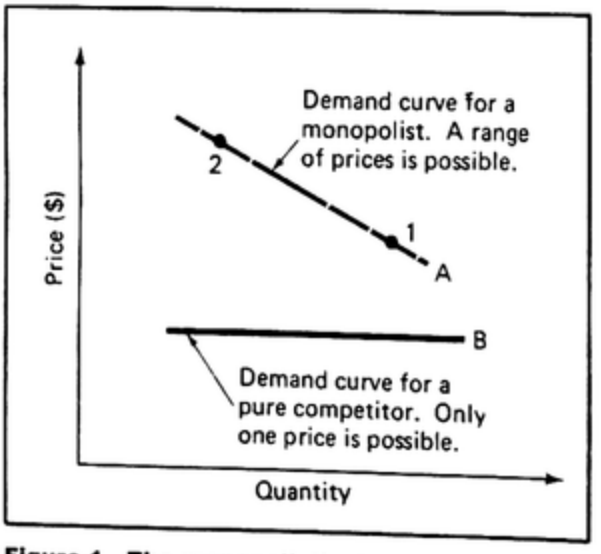


Figure 1 The monopolist's demand curve slopes down

The monopolist can set the price anywhere that it chooses along its demand curve. A pure competitor, by contrast, has no control over price. When the demand curve is horizontal, only one price is possible.

The monopolist's demand curve shown here is only a sample. Actual cases show great variety, from near-vertical curves to near-horizontal ones. Generally, the less elastic the curve, the higher the firm's degree of monopoly will be.

lower quantity and a higher price at Point 2. All other points along the demand curve are equally available. Meanwhile, the firm in a purely competitive market, with Demand Curve B, has no such degree of choice. If it raises its price even a little, it will sell nothing, for its customers will buy from other sellers. If a firm is a pure monopolist, controlling 100 percent of the market, then the market and firm demand schedules will be identical. If, instead, the firm has less than 100 percent of the market, its demand schedule will lie below or to the left of the market demand schedule. For example, in Figure 2, the market demand curve is shown along with the demand schedule of two of the firms in the industry. The firm represented by D_2 clearly has a larger share of the market than the firm represented by D_1.

Ideally, these curves would be known clearly and accurately, so that the degree of monopoly would be obvious. But de-

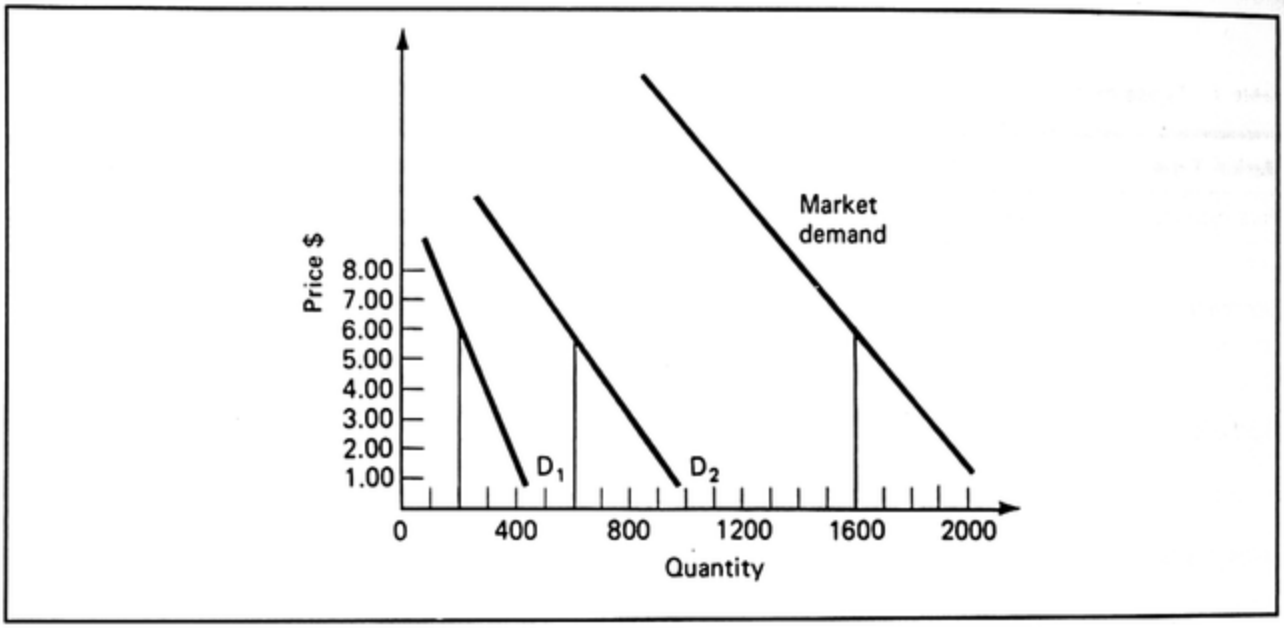


Figure 2 A comparison of firm and market demand

The demand schedules of two firms in this market are compared to the market demand schedule. The firm represented by D_2 has a larger market share, selling a higher quantity at each price than the firm represented by D_1. For example, at a price of $6, the total or market quantity demanded of the good will be 1,600 units. Of this total, 200 units will be demanded from Firm 1 and 600 units will be demanded from Firm 2. The remaining 800 units will be demanded from other firms in the industry.

mand curves aren't easily measured in real cases, as we noted in Chapter 5. So economists usually have to rely on other evidence to judge how much monopoly power a firm may hold. The main indicators are:

1. High market share A firm's ***market share*** is measured by its own sales, taken as a percentage of all sales in the market. "The market" is, in turn, defined to include goods that may easily be substituted for each other. A 100 percent share—the highest possible—is pure, total monopoly. A 10 percent share or lower usually gives the firm little market power. *Between 10 and 100 percent, the degree of monopoly power rises as the share rises.* A market share above 40 percent usually provides substantial monopoly power.

2. Lack of strong rivals If a firm is far larger than any other firm in its market, then its monopoly power cannot be strongly challenged. By contrast, the presence of equal or larger rivals would reduce the firm's ability to control the market.

3. Barriers to entry by new competitors Anything that makes it hard for new competitors to come into the market will enhance the market power of the firms already established there. If no ***entry barriers*** existed, even a firm with a 100 percent market share might conceivably have no market power.

4. High profitability Successful monopoly often leads to high profit rates on the firm's capital. Therefore, profit rates—taken together with the other signs of monopoly—often help the economist to evaluate how much market power a firm has. Yet, profitability is not conclusive evidence. Many monopolies fail to show high profits. They may hide their profits by accounting tricks; or they may become inefficient, make mistakes, or fall prey to other problems. Thus, high profit rates are not necessarily an indicator of monopoly power. Nor are they conclusive, since high

profits can also come from a company's sheer good luck, its recent innovations, or other causes. ***In short, high profit rates can help to confirm monopoly power, but they are not the key proof of its existence.***

In appraising real firms, economists try to consider these and other items. There is no exact formula, and the evidence about monopoly power will rarely be crystal clear. Utilities—with complete, officially protected monopolies—are usually straightforward cases. But the extent of the monopoly power of an IBM, General Motors, or Exxon is often hard to determine with precision. Like the best economists, you must rely on your judgment. Because most ordinary firms have little market power, as we will shortly see, economists concentrate mainly on dominant firms and tight oligopolies.

How monopoly power is created and maintained

There are many ways to gain monopoly power. Most firms try them most of the time, of course, but usually their efforts offset one another. Thus, competition itself tends to prevent monopoly. But when a firm does gain and hold a high degree of monopoly power, one or more of the following methods is usually at the root of it.

1. Mergers to capture a higher market share The firm may simply buy out its rivals, merging with them to get a high combined market share for the new, larger firm. Once unified, the former competitors no longer compete with one another.

2. Economies of scale Economies of scale are present when, because of the industry's technology, a large firm can produce a product more cheaply than a small firm. If the industry's cost curves show large economies of scale, then a firm will reach the lowest average cost at a high market share.

3. Superior innovation or efficiency A firm may capture most of a market by creating new and better products. Older products may be displaced, as when hand-held electronic calculators replaced desk-top electrical calculators in the 1970s. Or the new product may create a wholly new market, as Polaroid did with instant cameras in the 1950s and Xerox did with copiers in the 1960s. In either case, the high market share arises from innovation. Further innovations may help a firm keep its position, especially if its innovation is protected by patent.

Similarly, excellent management may allow the firm to gain and retain monopoly power. By cutting costs and inspiring workers, a firm's managers can cut prices and outsell their rivals. Being more ***X-efficient***—producing more outputs from given inputs—the firm earns its monopoly power and profits.

4. Official support Government policies often create or maintain monopoly power.

A. PATENTS are issued for inventions, giving a 17-year monopoly that the owner can exploit to the hilt. This can both provide high profits and be the basis for a lasting monopoly position. Crucial *trademarks*—such as Jell-O, Band-Aid, Kleenex, and Formica—can also have similar effects. They condition people to buy the product just because it has the familiar brand name. Such consumer loyalty allows the firm to charge higher prices without losing customers. In short, trademarks can make the firm's demand curve both higher and less elastic, thereby giving the firm a degree of market power.

B. MONOPOLY FRANCHISES are given to utility companies, excluding all others from competing. Your local electric, gas, telephone, and cable TV companies, for example, are protected by their government franchises. No other firms are permitted to

sell those services in your area. Taxi firms, banks, TV stations, and professional sports teams also usually have local franchise protection.

C. GOVERNMENT CONTRACTS often confer market power by making one or two firms the exclusive suppliers for large amounts of specific weapons (e.g., tanks, aircraft, cannon). Most military buying of weapons, for example, involves little competitive bidding among the aspiring suppliers of the armaments.

D. OTHER official supports include the Wagner Act of 1935, which permits labor unions to hold and exert monopoly power in labor markets. Many public enterprises—public schools and the U.S. Postal Service, for example—are also given monopoly power in their markets.

5. Key Inputs Some industries rely on key inputs, such as the ores that are crucial for metal industries. By controlling the input, a firm can monopolize all or part of the industry and keep new competitors out. Thus, steel and copper companies have sewn up most of the cheap ore supplies, making it difficult for new competition to enter their markets.

6. Unfair competitive tactics If a firm resorts to unfair methods, it may be able to drive out its competitors and gain monopoly power. "Unfair" action is not a clear-cut category. What's fair in love, war, and rugged competition is often debatable. But firms do often overstep the bounds of vigorous competition in ways that victimize and destroy their rivals.

Altogether, monopoly power has many possible roots. Some of them are praiseworthy, such as economies of scale and innovations. But others have no value at all: mergers, government favors, and unfair tactics, for example. The monopolist, of course, always claims that good reasons account for its position. Yet, less praiseworthy reasons for monopoly power may also—or instead—really apply.

Table 2 Features and sources of monopoly power

I. Features of Monopoly Power
1. A high market share, especially above 40 percent.
2. A lack of strong rivals with similar market shares.
3. Barriers to new competition.
4. High profit rates (a supplementary indicator).

II. Ways to Acquire Monopoly Power
1. Merging with competitors, and other actions simply to increase market share.
2. Achieving economies of scale (especially in "natural monopoly" situations).
3. Innovating more rapidly or managing production more efficiently than competitors do.
4. Getting official support for market control from:
   a. Patents on crucial products or techniques.
   b. Monopoly franchises, such as for utilities.
   c. Exclusive government contracts, such as for military weapons.
   d. Others.
5. Controlling a key input, such as a scarce ore or superior location.
6. Unfair competitive actions to harm rival firms.

Table 2 summarizes the main features and sources of monopoly power. Now we turn to the *effects* of monopoly power, whatever its origins.

The effects of monopoly power

Monopoly can have strong effects on prices, quantities and allocation in the market, the distribution of wealth, innovation patterns, and other economic values—in short, on all of the competitive outcomes discussed in Chapter 9. Table 3 lists monopoly's main effects.

The monopolist's choices *A pure monopolist is the entire supply side of the market.* Therefore, the monopolist's demand curve is the market demand curve itself.

Figure 3 shows the effects of turning a competitive market into a pure monopoly.

Table 3 ***The main effects of monopoly***

I. Monopoly *harms* economic performance:

1. *Efficiency in resource use* is reduced by changes in output and price:
   a. X-inefficiency may occur, raising average costs.
   b. Misallocation may occur, eliminating consumer surplus
2. *Equity in distribution* is reduced by monopoly profits (price discrimination may enlarge those profits). Wealth and income are shifted from the many to the few.
3. *Technical progress* (invention and innovation) is probably reduced. It becomes optional to the monopolist, and perhaps unprofitable because it reduces the value of the monopolist's assets.
4. *Broader values* may be harmed:
   a. Freedom of choice is reduced.
   b. Democracy is undermined.
   c. Culture and society become more closed and rigid.

II. There may be offsetting *benefits* from monopoly:

1. *Economies of scale* in production and innovation may be achieved.
2. Large innovations may be made more rapidly.

III. The *net effects* may go either way, in general and in each case. They require careful study, not mere slogans or assertions.

At first (in Panel I), the market is competitive; its supply and demand curves intersect to give the equilibrium price and output at Point *A*. Each competitive firm has a horizontal demand curve, though the market demand curve has the slope as shown.

Then a monopolist unites all of the suppliers into one firm. Its demand curve is now identical to the *market* demand curve, sloping down, as shown in both Panels I and II. The monopolist can now choose both the *price* level and the output level: Each price now corresponds to a different level of output. The monopolist's profit-maximizing choice fits the same basic logic as the competitive firm's (or any other firm's): *It sets output at the level where its marginal revenue equals its marginal cost.*

Marginal revenue is crucial to the monopoly outcome. The monopolist's *marginal revenue* curve is now below its demand curve, as shown in Panel II. When

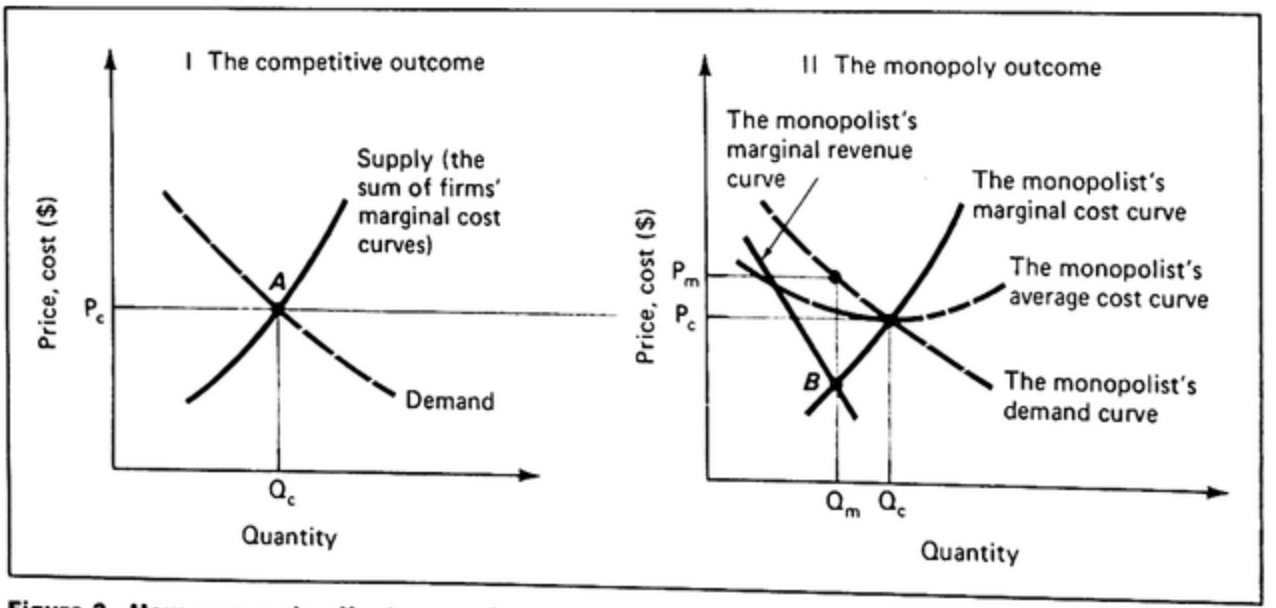


Figure 3 How monopoly affects quantity and price

Supply and demand set the competitive output and price at Point *A* in Panel I. If the market is monopolized by one firm, then the market demand curve is its demand curve. The monopolist now has a marginal revenue curve, as shown. The monopolist now maximizes its profit at Point *B* in Panel II, where its marginal revenue (*not* price) equals its marginal cost. At that output, Q_m, the price is P_m, which is above the competitive price P_c.

demand was horizontal for each competitor, marginal revenue was also horizontal and coincided with demand for each firm. Selling another unit brought in exactly as much revenue as the price itself. But now that there is just one firm, its demand slopes down. To sell more units, the monopolist must cut its price, and *that* reduces the price obtained on all of the units that would have been sold at the higher price.

For example, in Figure 4, the price must be cut from \$12 to \$11 to sell the eighth unit. That brings the firm an additional \$11 for selling that unit (Area *A*). But the price on units one through seven would have been \$12. Cutting to an \$11 price to sell an additional unit means that the firm gives up \$1 on each of those seven units (shown by Area *B*). The *net* gain from selling unit eight is thus \$11 minus \$7 (Area *A* minus Area *B*), which is only \$4. The marginal revenue of the eighth unit is therefore \$4, which is well below the \$11 price.

The marginal revenue from selling more output is the price of the added unit *minus* the revenue lost by cutting price on the other units. This value must be less than the price on the demand curve. Therefore, the marginal revenue curve always lies below the demand curve.

The marginal revenue curve is easy to locate: The marginal revenue curve of a straight-line demand curve is always a straight line, halfway between the demand curve and the vertical axis. You merely draw two light horizontal lines to the left of the demand curve, mark the halfway points (Points 1 and 2 in Figure 4), and draw the marginal revenue curve through them. Or, alternatively, you draw and divide one horizontal line, find the point on the vertical axis that the demand curve would go through, and then use these two points to draw the marginal revenue curve. With this skill, you can always place the marginal cost curve roughly correctly, in relation to the firm's demand curve.

Where a monopolist has unified all of the firms in a formerly competitive market, its marginal cost schedule will be the former competitive market's supply curve (as was shown in Figure 3). Remember that this supply curve was the sum of all the competitive firms' marginal cost curves. Since the single monopolist now includes all of those firms, *its marginal cost*

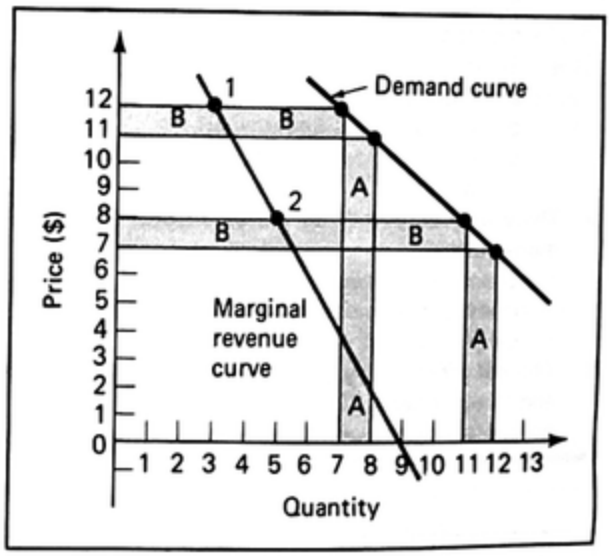


Figure 4 When demand slopes down, the marginal revenue curve lies below it

Marginal revenue shows the net effect on revenue of producing another unit. The revenue from an extra unit is added (Area *A*), but the cut in price for all the other units must be subtracted (Area *B*). The same net calculation can be done for any point along the demand curve. Thus, at an output level of 7, *A* is bigger than *B*, and marginal revenue is positive at a \$4 value. But when output is 12, *B* is bigger than *A*, and marginal revenue is negative.

The exact numbers for the two cases shown in the figure are:

Output	Price	Total Revenue	A	B	Marginal Revenue
7	12	84	+11	−7	4
8	11	88			
9	10	90			
10	9	90			
11	8				
11	8	88	+7	−11	−4
12	7	84			

curve is the summation of all the original firms' marginal cost curves.

The monopolist's two curves—marginal revenue and marginal cost—cross at Point *B* in Figure 3, which is thus the monopolist's profit-maximizing output. As always, the firm produces only up to the level at which the extra unit is just "worth what it costs to produce it." That is the commonsense meaning of marginal revenue just equaling marginal cost.

As Figure 3 shows, the monopolist has chosen output level Q_m. The price that results is shown by the point on the demand curve at output Q_m. Consumers will be willing to pay P_m (*m* stands for "monopoly"), so that becomes the monopolists' profit-maximizing price. One might say that the monopolist chooses the output rather than the price; but, in effect, *the monopolist's decisions set both output and price, whereas competitors merely take the price as given.*

Recall also from Chapter 9 that the competitive outcome made prices equal to marginal costs. Price is a measure of the social value of a good, for it shows what people are willing to pay for it. Marginal cost is the true cost—the opportunity cost, the social sacrifice in terms of the resources used—of producing a specific amount of the good. The $P = MC$ equality means that the output of the good is expanded up to the point at which the cost of the resources used to produce the last unit of the good just equals the value that the consumers place on that last unit. This is the general condition of efficiency, since only at this point is the social value of the good equal to its cost.

The monopolist violates that efficient outcome by cutting output and pushing price above marginal cost. In Figure 5 the new monopoly price is much higher than marginal cost. There is economic harm in that disparity. People are willing to pay

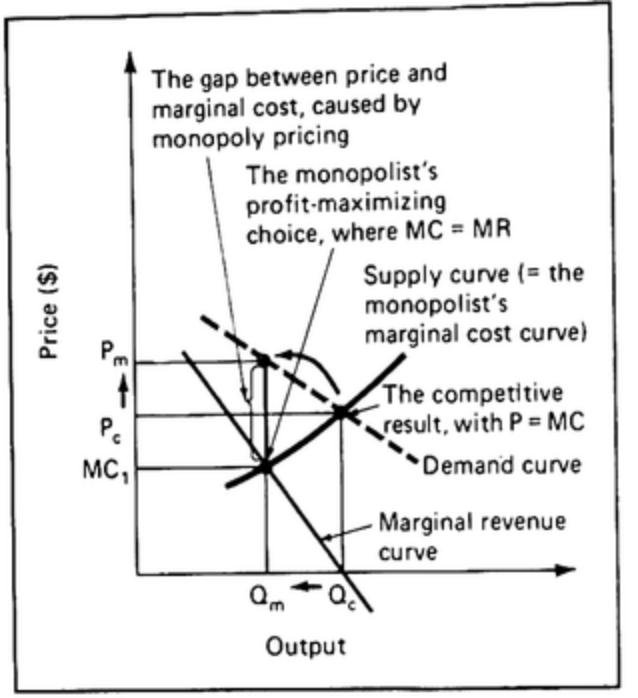


Figure 5 The simple effect of monopoly

The efficient competitive result is where supply equals demand, with Q_c and P_c (c stands for "competitive"). The new monopolist now interferes. It heeds its own marginal revenue curve, setting its output level where marginal revenue equals marginal cost. Output is cut to Q_m and price rises to P_m (*m* stands for "monopoly"). Price is now well above marginal cost, by the distance $P_m - MC_1$. Consumers would pay up to the price P_m to get output that costs only MC_1 to make, but the monopolist won't let them do that. Therefore, output and prices are distorted from efficient levels. (Note: The marginal cost schedule is *not* the monopolist's supply schedule, as you will see.)

more than twice as much as the marginal cost of the good. The value that they place on the good is twice as high as the cost of the resources used to produce the last unit. This is a clear signal that more of the good should be produced. But the monopolist has cut output back, effectively prohibiting sales between Q_m and Q_c. Therefore, it can force people to pay a monopoly price that is well above the real cost of supply. ***Price exceeding marginal cost is, therefore, a sign of distortion away from the efficient competitive level. It calls for an expansion of output, but the monopolist prevents that expansion from occurring.***

The severity of this cutback in output depends on the elasticity of demand and the slope of the monopolist's marginal cost

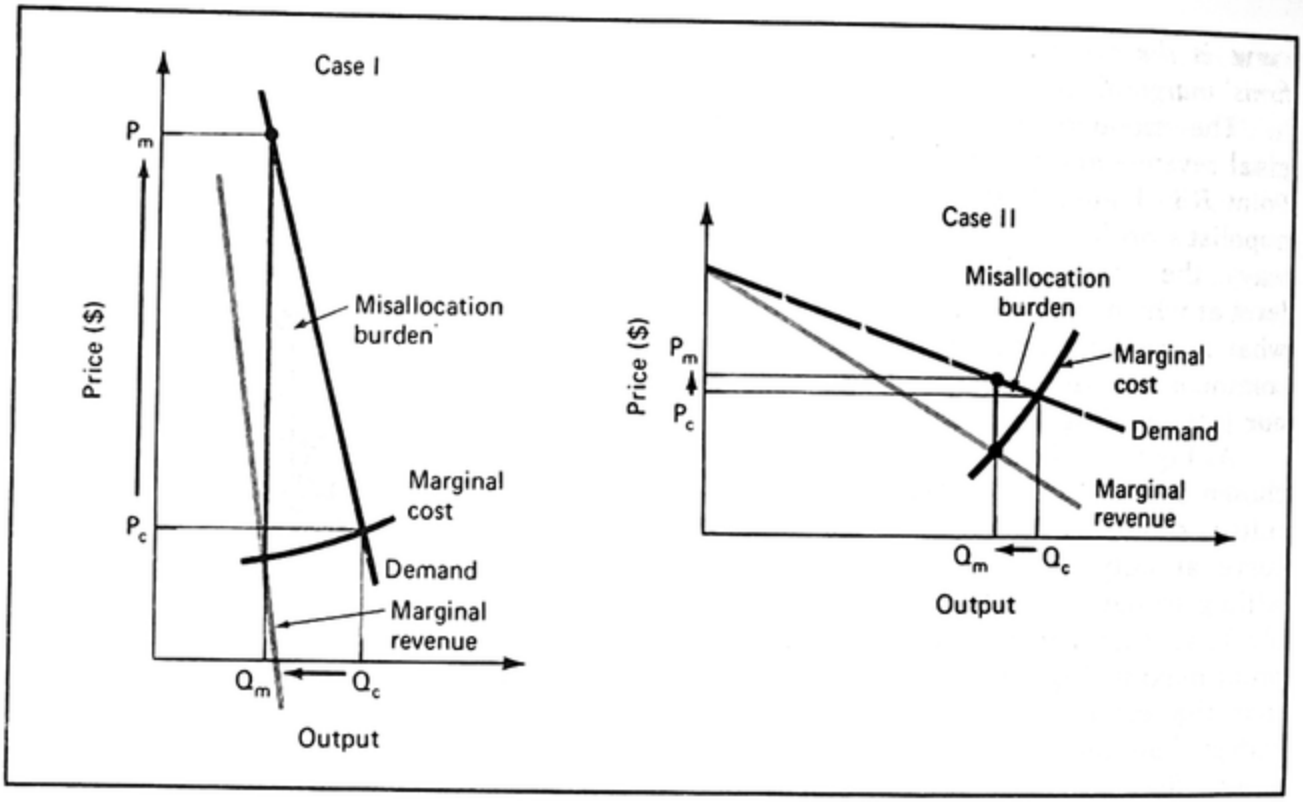


Figure 6 The severity of monopoly's effects depends on demand and cost conditio ıs

In Case 1, because demand is extremely inelastic, consumers can be sharply exploited. The monopolist's sharp cut in output causes price to multiply.

In Case 2, because buyers apparently have good alternatives to the monopolized product, their demand is more elastic. Also, marginal cost is steep, so that output is not cut back by much. Since both price and output are scarcely changed, monopoly has little effect on allocation.

curve. Compared to elastic demand, inelastic demand gives smaller cuts in output but sharper rises in price. As for cost, the steeper the marginal cost curve, the smaller will be the difference between the monopoly price and marginal cost.

Figure 6 illustrates these conditions. In Case 1, demand is inelastic and marginal cost is relatively flat. Because of these two conditions, monopoly has drastically changed both output and price. The price is four times the original price and five times marginal cost. Output is only about half of the competitive level. When people urgently need an item (a necessity, for example), a monopolist can severely exploit them. In contrast, Case 2 shows only a mild effect: Demand is more elastic, and the marginal cost curve is steeply sloped. In this case, price is only nudged up—and output down—a little.

A pure monopoly may have a severe, slight, or moderate effect, all with the same 100 percent market share. These are matters of *degree*, depending on conditions. The severity of the effects can be predicted if one knows the underlying conditions of demand and cost. The *logic* of the effects of monopoly is the same in every case.

THE MONOPOLIST HAS NO SUPPLY CURVE

Recall that the purely competitive firm faces a horizontal demand curve at the going market price. As that demand curve shifts up or down in response to changes in market price, the firm sets its output at the level where demand intersects the

marginal cost curve. The competitive firm's marginal cost curve is therefore also its supply curve.

The monopolist differs. Its demand curve slopes down and is separate from its marginal revenue curve. Moreover, the demand curve can shift in any way—rotating, parallel, bending—rather than be rigidly horizontal. Shifts in the demand curve, therefore, do not trace out a supply curve that coincides with the marginal cost curve. In fact, they do not trace out any supply curve at all.

To verify this conclusion, you can try a series of demand-marginal revenue shifts in a diagram, finding the new profit-maximizing price and quantity combination in each case. You will see that the points are much too scattered to be on any single supply curve. Consider Figure 7. The original demand and marginal revenue schedule, D_1 and MR_1, result in a profit-maximizing price and quantity combination of P_1 and Q_1. Therefore, the firm would want to supply Q_1 at a price of P_1, given the present demand and supply conditions.

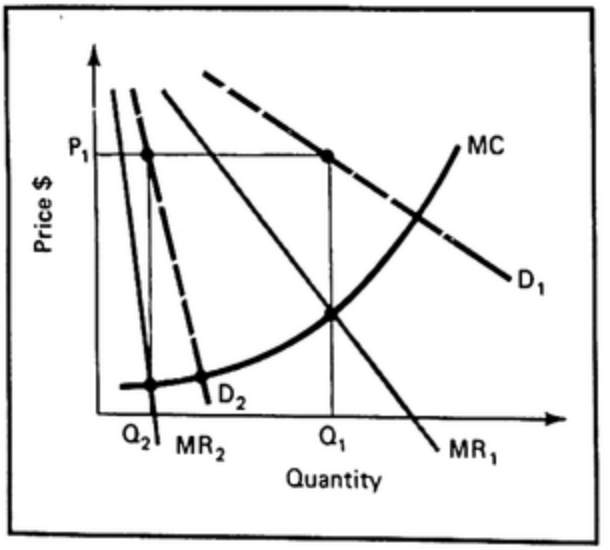


Figure 7 The lack of a unique quantity supplied at a given price

With demand and marginal revenue represented by D_1 and MR_1, the firm will supply the profit-maximizing quantity of Q_1 at a price of P_1. Notice what happens when demand and marginal revenue shift to D_2 and MR_2. The firm will now find it profit maximizing to supply Q_2 at a price of P_1. Given different demand conditions, then, different quantities may be supplied at a price of P_1. Therefore, a schedule representing a distinct quantity for a given price cannot be drawn.

Suppose that the demand schedule and therefore the marginal revenue schedule shifted to the left, as shown by D_2 and MR_2. Now the profit-maximizing price-quantity combination results in a lower quantity of Q_2, but the price remains at P_1. Therefore, under different demand conditions, the firm may supply different quantities, Q_1 or Q_2, at a price of P_1. A supply schedule showing unique price-quantity combinations cannot be drawn.

Note that *the monopolist does not raise prices as high as it possibly can.* Instead, it raises prices just to the level that maximizes profits. That rise may be small or big, as Figure 7 illustrates. But it is not the maximum possible price. The maximum price would be at the left end of the demand curve, where the monopolist would sell just one unit.

Monopoly's effects on economic performance There are also other effects of monopoly, such as misallocation, unfair redistribution, X-inefficiency, and a possible slowdown on invention and innovation in a given field. We will now consider these effects one by one, starting with the shift away from the competitive allocation.

MISALLOCATION Because it reduces output, monopoly distorts the allocation of resources. Output is held below the level at which price equals marginal cost. The contrast is shown by Q_m and Q_c in Figures 3, 5, and 6. The cutback in output forces some of the inputs into other markets, where their economic value is less. These distortions ripple through adjacent markets into the whole economy. Monopoly in one part of the economy disturbs the functioning of the whole system. The larger the

monopolized industry is, and the more severe the direct effects are (as in Case 1 as opposed to Case 2 in Figure 6), then the greater the economic harm will be.

The distortion is called *misallocation.* It is caused by moving resource use away from the efficient pattern. There is a loss of economic value, a loss that shows up as the reduction in ***consumers' surplus. By forcing price down into line with cost, competition maximizes consumers' surplus,*** as shown in panel I of Figure 8. Monopoly's effect can now be seen clearly. By raising the price, as in Panel II of Figure 8, the monopolist eliminates some of this consumers' surplus. Compared to perfect competition, the monopolized market outcome, determined by the marginal revenue–marginal cost intersection, will be a lower output (Q_m instead of Q_c) and higher price (P_m instead of P_c). Note how the consumers' surplus shrinks from *ABC* to *ADE* as market price increases. The total loss in consumers' surplus is made up of two components. The rectangle *EDFC* represents the increase in payments by consumers, in the form of excess profits. It is a redistribution of income from consumers to producer because of the higher price. The remaining portion of the reduction in consumers' surplus, the small triangle *DBF*, represents the welfare loss to society resulting from the resource misallocation that monopoly causes. It is the value that consumers placed on the output that is lost as a result of the monopoly.

Economic welfare is reduced: *The monopolist destroys the economic value shown by the triangular-shaped area of consumers' surplus, by changing the allocation of resources.* Therefore, the burden of misallocation caused by the monopoly is shown

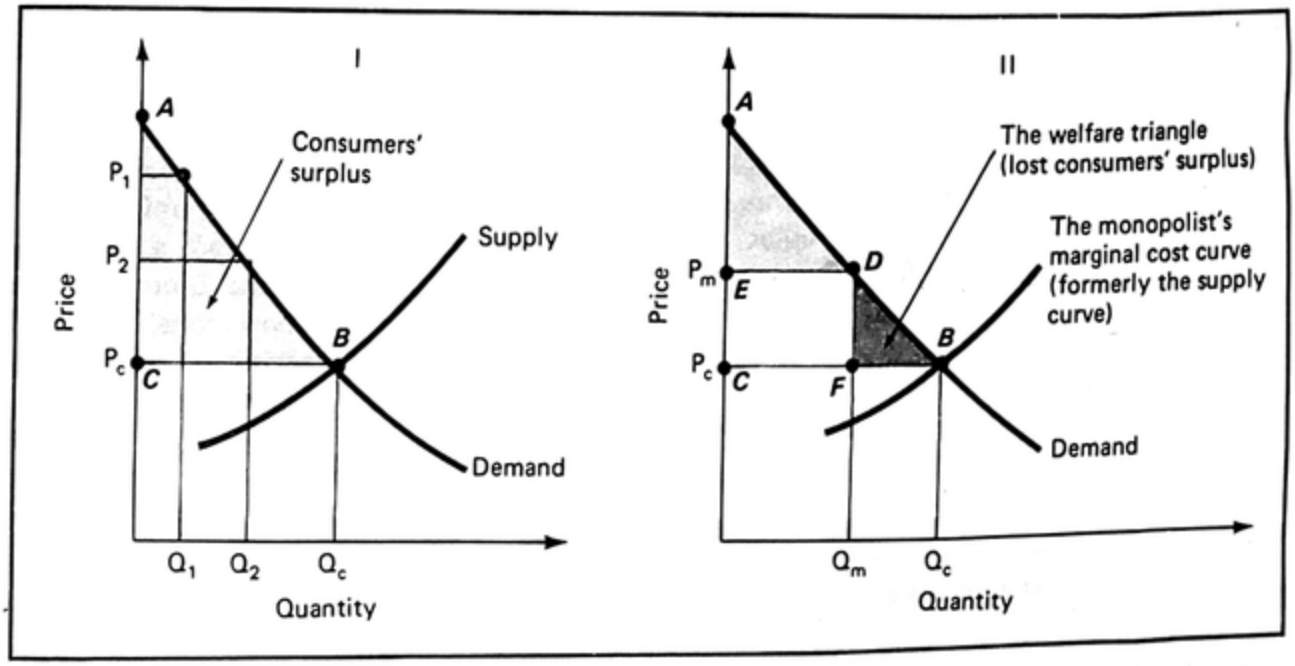


Figure 8 Consumers' surplus and monopoly's destruction of it

Consumers' surplus under competitive conditions is shown by the triangle *ABC* in Panel I. It stands for the difference between the value that consumers place on a commodity, represented by points along the demand schedule, and the market price. Since price under competition equals minimum average total cost, consumers' surplus is maximized under perfect competition.

Monopoly will result in a higher price and lower output, as Panel II shows. Consumers' surplus shrinks from *ABC* to *ADE*. The reduction in consumers' surplus has two components: the redistribution from consumers to producer (the rectangle *EDFC*) and the loss in output *(DBF).*

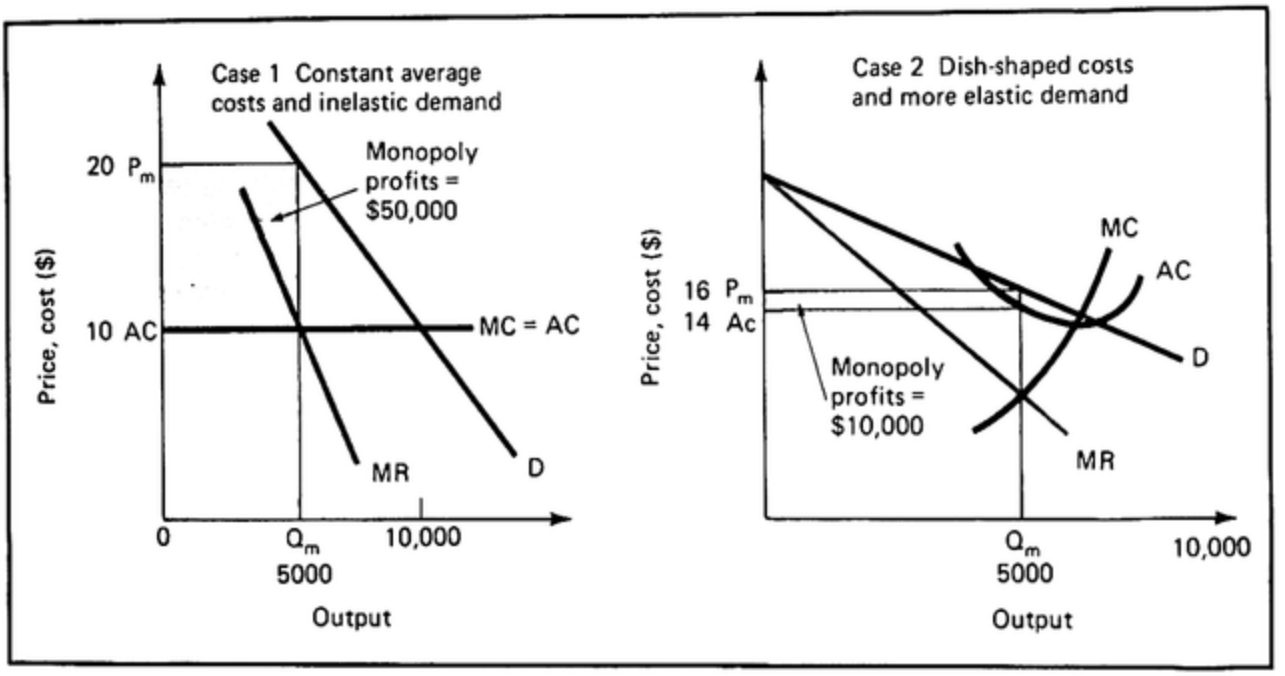


Figure 9 The monopolists' profits may be large or small

Excess profit is the gap between the demand curve (which is the average revenue schedule) and the average cost curve (that is, excess profit per unit) *times* the number of units sold. In Case 1, excess profit is $20 minus $10 = $10 per unit, *times* 5,000 units = $50,000. In Case 2, excess profit is $16 minus $14 = $2 per unit, times 5,000 = $10,000. The shaded areas show the excess profits. These profit volumes are the areas of the rectangles.

by this ***"welfare triangle,"*** as economists call it.

The ***misallocation burden*** may be large or small, as illustrated in Figure 6. In Case 1, the burden is large, 40 percent of the monopolist's total sales revenue. In Case 2, the burden is small, perhaps just 1 percent of sales revenue.

UNFAIR REDISTRIBUTION The monopoly may also provide *monopoly profits*, in excess of the normal profits gained by competitive firms. The amount of profits is set by the gap between price and average cost, at the monopolist's level of output, as Figure 9 shows. Total excess profits are calculated by multiplying profit per unit times the number of units sold. In Figure 9, that is P_m-$AC \times Q_m$. The magnitude of the excess profits depends on the positions and shapes of the demand and cost curves. In Case 1 in Figure 9, because the steep demand curve is well above average cost at the monopoly output, the excess profits are large. But in Case 2, demand is down close to average cost, so that this monopolist gains only a small excess profit. You can draw other cases, illustrating medium or even zero monopoly profits.

The excess profits—whatever their level—usually represent a degree of unfairness. They redistribute income, transferring money from the pockets of ordinary consumers into the monopolist's till. Many consumers will lose income, while the one (or few) owners of the monopoly will gain sharply. The consumers usually have lower incomes than the monopoly's owners.

Therefore, monopoly tends to tilt the income distribution toward greater inequality. These income flows are capitalized into wealth, as we explained in Chapter 7. A monopoly's stock price will rise to reflect the flow of monopoly profits, and

the owners can sell out immediately and put their wealth into other investments. Thus, monopoly creates family fortunes, enriching a few at the expense of many. This shift of wealth is usually quite unfair, since the buyers are ordinary people who can ill afford to pay higher prices to enrich a few monopolists.

X-INEFFICIENCY There is another way in which monopoly may reduce efficiency. Being free from the pressures of competition, the monopoly firm's management may lose some of its tightness and vigilance. Cost controls may not be as strict and productivity may decline because everyone working in the firm knows that the firm is profitable and that it won't go out of business if costs rise.

This internal slack is X-inefficiency. It differs from *allocative* inefficiency among firms and markets (which the welfare triangle shows). *In a monopoly firm, X-inefficiency may cause a simple rise of the cost curves above their lowest possible levels.* Such X-inefficiency can range from small to large.

These effects are defined in a *static* context. (Compare them to the competitive equilibrium in Table 2's summary.) Monopoly also has effects in a *dynamic* context, altering the rate of invention and innovation.

INVENTION AND INNOVATION A monopoly is usually not under pressure to *invent* new products or methods. Nor does it have strong incentives to *innovate:* to apply those new inventions in practice and bring new products to the market. *The monopoly may choose to invent and innovate, but it will do so only at its own pace.* No competitor forces its hand. Even if its capital is outdated or its products mediocre, a monopolist may prefer to protect and continue them rather than to replace them with better ones.

That monopolies often retard innovation has been common knowledge for centuries. The problem has been studied closely and two main lessons have emerged: (1) Monopolists may *invent actively*, so as to know which new ideas are coming. But (2) they usually *innovate* more slowly than competitive firms would. In short, a monopoly usually retards progress.

There may be certain exceptions, however. An innovation may be so big, costly, and risky that a monopolist—with its large size and high degree of security in its market—is better able than a hard-pressed competitive firm to carry it out. Large companies sometimes suggest that such cases are frequent. That is a matter on which judgments vary. Yet note: Even when an innovation does require vast resources to carry out, the monopolist's tendency to retard innovation will still be at work. In short, the monopolist may be able but not very willing.

Another exception can occur when the monopolist is insecure, vulnerable to being dislodged by an innovator. In this case, the monopolist may innovate swiftly because of the fear that the newcomer may do the innovation instead. Yet, the insecurity must be genuine, which may be rare for well-established monopolies. A secure monopolist usually retards progress.

OTHER EFFECTS OF MONOPOLY Monopoly restricts *freedom of choice* for everyone involved except the monopolist. Buyers cannot try other suppliers; they are stuck with this one monopolist, for good or ill. Only those goods that the monopolist offers are available in this market, and they carry higher prices to boot. The former competitors are out of business, or working for the monopolist. Newcomers may be unable to enter the market. If entry barriers are high, only the strongest entrants may have a chance; or perhaps no entrant can sur-

vive, so that the choices of former and potential competitors are reduced.

Suppliers also have less choice. Not only are their sales cut back by the monopolist, but they also have less opportunity to offer new products to a variety of firms—there is only the monopolist to sell to in this market. Workers also have fewer choices. They must deal with only the one monopolist in the industry. Everyone—buyers, suppliers, workers, and would-be-competitors—loses freedom of choice.

Democracy is also affected by monopoly. There are fewer firms with less diversity of interests. The monopolist is now a power bloc, with a valuable advantage—and perhaps excess profits and market power—to protect. By supporting friendly candidates, by seeking favorable laws and rulings, and by advertising its interests via the media, a monopoly can use the political process to protect and enlarge its economic position. Even when its actions are mainly subtle, monopoly is likely to undermine democracy.

Culture and society can also be affected. When many markets are monopolized, the economic and social order becomes tight and closed. Society is more stratified and rigid, less open to outsiders and new ideas. Fascism, for example, grew partly out of societies that had market power in many key parts of the economy. In another vein, monopolists can often influence consumers' preferences without challenge from others. An economy of monopolies provides a distinctive and unattractive social content, going against many traditional American values.

Price discrimination

Another special effect of monopoly is ***price discrimination.*** Its precise definition is: *different price-cost ratios to different customers, rather than one price-cost ratio to all.* In plain English, discrimination means "charging what the traffic will bear" in each part of the market: "selective pricing," "pricing by market segmentation"—in short, price differences based on the differences in consumers' *demand.*

Monopolists use price discrimination to extract more profit, to improve their market positions, and to defeat their competitors. But all firms try to discriminate—to price selectively *where they can*—and the practice in itself is neither good nor bad. Consider now what it is, how widespread it is, and when it is harmful or helpful to competition and the economy.

Preconditions for discrimination Price discrimination can occur when three conditions hold:

1. Buyers have sharply differing demand elasticities.
2. The seller knows these differences and can separate the buyers into groups based on these differing elasticities.
3. The seller can keep the buyers from reselling the product or service to one another.

Under these conditions, the seller will divide the buyers into two or more groups and then charge higher prices to the buyers who have the less elastic demand. Remember that inelastic demand means a higher degree of urgency or need. Those who would pay more *are made* to pay more. Other buyers with more elastic demand—who have good substitutes or simply can't afford to pay more—are charged less.

The classic instance has been the town doctor who treats all comers, rich and poor. For the same appendectomy, the banker is asked to pay $800, the poor widow $50. Nineteenth-century railroads were also masters at charging what the traffic would bear: typically, 10 cents per ton-mile out on the plains, and 2 cents per

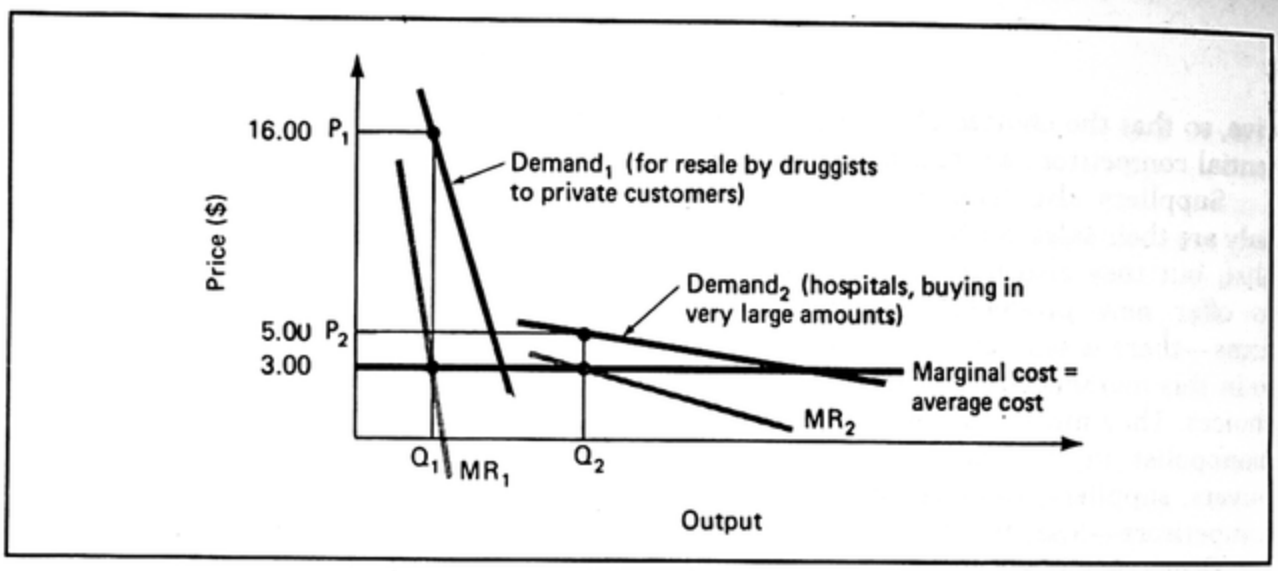


Figure 10 Simple price discrimination

The same drug (with uniform costs) is sold to two groups: one is druggists for resale. Demand is inelastic because buyers merely take what their doctors prescribe. Those sales are priced at $16 per dozen, to maximize profits. Hospitals (Group 2) can shop around and drive hard bargains. Therefore, large-volume sales to hospitals are at $5 per dozen, again maximizing profit on that part of their sales.

If hospitals open drug shops to resell the drug, at any price between $5 and $16, this price discrimination would weaken and perhaps disappear. Would drug firms and druggists oppose this step, by arguments and lobbying for laws to prevent such "unethical" or "hazardous" practices?

ton-mile alongside rivers with competing barge lines. Perhaps the most familiar instance today is half-price movie, bus, train, and airplane tickets for children. The costs are much the same for both children and adults, yet adults pay a much higher ratio of price to cost.

When discrimination occurs, the elasticity of demand—*not* cost—governs prices. A price discriminator will follow the same basic rule as any other profit-maximizing firm: Set the price and output at the level for which marginal revenue equals marginal cost. A single-price firm will be working with the demand and marginal revenue schedules for the entire market. A price discriminator, however, will set price for each group of customers on the basis of the demand and marginal revenue for that particular group.

Figure 10 shows a typical case of price discrimination, using a lifesaving drug as an example. The same drug is sold to two groups: (1) to druggists for resale to individuals, by doctors' prescriptions; and (2) to large hospitals, for dispensing to patients. The druggists' customers have *low* demand elasticity, for they merely buy what their doctor writes on their prescriptions.

By contrast, the hospitals have *elastic* demand. They can bargain shrewdly, playing off the drug companies against one another to get a low price. Thus, the identical drug, costing perhaps $3 per dozen pills to make, might sell for $16 per dozen to retail druggists and $5 per dozen to hospitals. (In practice, the ratios of price to cost often differ even more sharply.) The inelastic demand results in a higher price to one group, which in this case is the retail druggists.

Many other familiar situations give rise to price discrimination, some of which are listed in Table 4. The critical fact is the differing price-cost ratios among customers. In the drug instance:

$$\frac{\text{Price}_1}{\text{Cost}_1} = \frac{\$16}{\$3} = 5.33 \quad \text{which does not equal} \quad 1.67 = \frac{\$5}{\$3} = \frac{\text{Price}_2}{\text{Cost}_2}.$$

Table 4 ***Instances of price discrimination—and not discrimination***

Good or Service	Consumer Groups	Costs and Prices	How Reselling Is Prevented	How Does This Affect Competition?
1. *Prices differ more than costs: DISCRIMINATION*				
Movies, airplane trips, train trips	Adults, children	Costs are about the same for all customers, but children pay much less than adults	By letting only children use children's tickets	Usually not much at all
Magazines	Newsstand sales, regular subscriptions, special subscriptions, (to new subscribers, students, etc.)	Costs are about the same, but prices differ sharply among customers	Magazines are bulky and easily damaged, subscriber lists are kept separate	Often it promotes competition; rarely does one magazine dominate the market
2. *Costs differ more than prices: DISCRIMINATION*				
Bus trips in town	People taking various-length trips	There are uniform fares, but costs differ for different lengths	Tickets are issued only for the ride	Not much for the most part because competition is already excluded by the bus franchise
Electricity	Various times of day and week	Prices per kilowatt-hour do not vary by the time of use, but peak-time costs sharply exceed off-peak costs	Storing electric energy is difficult; reselling is illegal	Not much because competition is already excluded by the utility's franchise
3. *Prices and costs differ proportionally: NOT DISCRIMINATION*				
Long-distance phone calls	Daytime callers, night and weekend callers	There are higher prices for calls made at peak times; costs at peak times are also higher	Timing cannot be switched	Not much, for the telephone company has an exclusive franchise
Clothing sales	Regular sales, bargain sales	There are low sales prices; the true costs of the clothes are also low, because clothes are excess (recall opportunity costs)	Sales are held only after clothes have lost their popularity	It promotes competition, for the firms rarely have large market shares
Restaurant meals	Luncheon customers, supper and evening customers	At peak-capacity times, costs are higher, even though food costs are uniform; dinner prices are well above luncheon prices	By time of day; meals cannot be stored or resold by those who buy them	It promotes competition by filling restaurant tables at flexible prices, rarely do restaurants dominate their markets

To make price discrimination work, the seller must keep the low-price buyers from reselling the product to the high-price buyers, for such reselling would pull the high price down toward the low price. In all of Table 4's instances, the seller has special ways of preventing the customers from reselling to one another.

Note that uniform prices can be discriminatory if costs differ. For example, the price is 20 cents to mail a first-class letter anywhere in the United States, across town or from Maine to California. Because that uniform price ignores the greater costs of the longer routes, it is discriminatory (though not necessarily bad, and possibly quite good). One judges possible discrimination by comparing *price-cost ratios*, not just prices. Cost differences can justify price differences.

Note, too, that many cases of discrimination are neutral or actually procompetitive, rather than a threat to competition. In fact, many little firms—which hold no market power—compete by selective price cutting. Such price cuts, often called "loss leaders," are common in grocery, drug, clothing, and camera stores. Several items are temporarily offered at discounts, to draw customers in. Once there, the customers may buy other goods that have higher profit margins. Newspaper ads for grocery and clothing sales are often full of such "loss leaders."

Discrimination can be the lifeblood of competition when it is *sporadic* and/or done by *smaller* firms. ***Only when it is done forcefully and systematically by dominant firms is price discrimination usually anticompetitive.*** That is shown by Figure 11. Indeed, if firms with small market shares try to keep prices systematically out of line with costs, they will lose money and may go out of business. At the other extreme is the utility firm, holding a complete monopoly and selling to many different buyers. It is always tempted to apply deep price discrimination. That is one reason why utilities in electricity, gas, and telephones need to be regulated by public agencies.

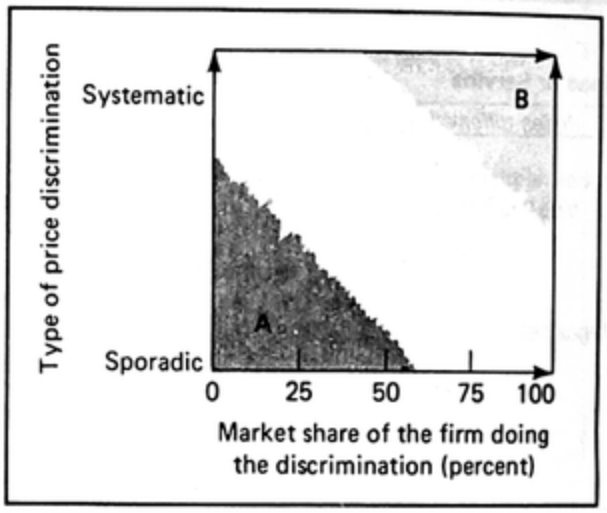


Figure 11 Price discrimination can be procompetitive or anticompetitive

In Area *A*, the discrimination is done sporadically by a firm with a small market share. The result is flexible pricing, which may increase competition. In Area *B*, the firm has a large market share and it practices systematic discrimination. That increases its excess profits and reduces the ability of lesser firms to compete.

You can discover many cases of price differences in familiar markets. For each case, test (1) whether costs also differ by the same proportion, and (2) whether the discriminator is a dominant firm or a little competitor. Often, systematic price discrimination is a sign that the firm holds market power. Only a firm with a high market share can maintain systematic discrimination in its prices.

Cases of monopoly

Knowing what monopolists are and do, you can now examine some case studies of real-life monopoly power. Though not all of them attained pure monopoly, they illustrate the effects of high degrees of monopoly power. When economists try to

evaluate monopoly and its effects, they look especially for raised prices and excess profits, for price discrimination and for X-inefficiency.

Standard Oil

America's most spectacular monopoly has probably been the Standard Oil combine. It was formed in the 1870s by John D. Rockefeller, a relentless, hard-bargaining, tightfisted, puritanical man. It became the Standard Oil Trust, combining a series of Standard Oil companies that were monopolies in their states and regions (thus, Standard Oil of New York, of New Jersey, of Ohio, etc.). For nearly 40 years, this industrial combination controlled between 60 and 90 percent of U.S. oil production—nearly a complete monopoly—and it yielded large monopoly profits. Then, in 1911, a climactic antitrust decision by the U.S. Supreme Court divided Standard Oil into its 33 parts, many of which were regional monopolies.

By the 1930s, competition had set in throughout the oil industry, though the successor oil firms have often tried to avoid open price competition. (The present descendants include Mobil—formerly Socony-Mobil, the Socony being *Standard Oil Co. of New York*; Standard Oil of Ohio; of Indiana; of California; and other firms. Standard Oil of New Jersey was the main successor. It changed its name to the anonymous-sounding "Exxon Corporation" in 1971.)

Standard Oil probably raised oil prices by over 30 percent on average, after driving out or merging with its competitors. Its profit rates were over 60 percent on capital before 1911. That is an exceedingly high rate, as Chapter 7 showed. The flow of monopoly profits totaled over $1 billion by 1911, creating an immense Rockefeller family fortune, which was soon applied in other industries. The profits were extracted from competitors and customers, shrinking their wealth.

Standard Oil used price discrimination to weaken its small rivals. It cut prices selectively in one area after another, often forcing single small competitors to go out of business or to sell out to Standard Oil at reduced values. In many cases, it needed only to threaten to cut prices to get its way. Though some economists deny that Standard Oil actually killed off small rivals by using price discrimination, the selective pricing certainly helped reduce their ability to compete.

Opportunities for others to do business in the industry were sharply curtailed. Standard Oil also corrupted legislators, railroads, and other businesses. Indeed, some of its growth came from forcing the railroads to pay Standard a sum of money for every barrel of *competitors'* oil that the railroads shipped!

Standard Oil strongly displayed most of the classic features of monopoly. It was widely hated and feared, and the final antitrust action came after a nationwide groundswell of discontent and legal attacks. The modern distrust of the oil industry has deep roots in the Standard Oil Trust.

Electric companies

In the decades after Thomas Edison invented the electric lamp and the technology of electric distribution in 1879, electric companies grew and spread from local to regional systems. There was an early period of competition in many cities, but most electric companies soon merged and gained exclusive franchises in their service areas. These combinations mainly reflected large economies of scale—the downslope of average cost curves—that gave "natural monopoly" conditions (one set of wires per town is much cheaper than two). Until these companies eventually

were regulated in the 1920s and 1930s, they behaved as monopolies generally do. Many of them set price levels to earn high rates of return. Price discrimination was rampant, tailored to customer groups' demand elasticities. The management of electric utility companies often became slack, sheltered as it was by monopoly.

During 1910–1940, as electric systems grew and broadened, regulation also spread. It gradually reduced the utility firms' profit rates toward the cost of capital and limited the more obvious kinds of discrimination. But some price discrimination persists even today in the 170 private electric systems and in various public ones, too. For example, electricity prices are usually uniform rather than varied to fit differing costs at peak times during the day and off-peak times at night and on weekends. Management also retained a degree of X-inefficiency in many electric firms up through the 1960s. Yet, since 1965, pressures from rising oil prices and new technology have forced utility companies to become more efficient. Therefore, on the whole, electricity firms have exhibited some of the normal monopoly behavior, even though regulation is supposed to prevent it.

Summary

1. Monopoly power is an ancient and continuing problem.
2. Pure monopoly and pure competition are both unusual. Most markets have a mixture of monopoly and competition, ranging from dominant firms down to monopolistic competition. Each market's degree of competition is usually a matter of debate.
3. Monopoly usually hurts economic performance. The monopoly restrains trade in the market by reducing output and raising price. This causes the allocation of resources to be distorted, as shown by the welfare triangle.
4. Monopoly may also encourage slack management and slow innovation. It enriches a few at the expense of many. It shrinks freedom of choice, and its concentration of power can undermine healthy democracy.
5. There may be economies of scale or large innovation that offer social benefits.
6. Natural-monopoly conditions resulting from substantial economies of scale may make competition impossible. Yet, the economic harms of monopoly still occur even when there are gains from reducing costs. The economic task is to compare the harms with the possible gains. Often a firm will gain much more monopoly power than scale economies can justify. In practice, monopoly's effects have occurred in many industries, often mildly but occasionally severely.
7. Price discrimination, which causes differences in price-cost ratios, is often practiced sharply by monopolists. When it is done systematically by dominant firms, it reduces competition. But done sporadically by lesser competitors—as, in fact, it frequently is—price discrimination can promote competition.

Key concepts

Monopoly power or market power
Market share
Entry barriers
Natural monopoly
X-efficient and X-inefficient
Consumers' surplus
Welfare triangle

Misallocation burden
Price discrimination

Questions for review

1. a. Consider the following list of firms. On the basis of your own knowledge of firm and industry characteristics, classify each according to the market types found in Table 1. If possible, compare your classifications with those of other students in the class.
   i. General Motors
   ii. A & P
   iii. Wheat farmer
   iv. U.S. Steel
   v. Local clothing store
   vi. Procter & Gamble
   vii. Time-Life Publishing Company
   viii. Michigan Consolidated Gas
   b. Now try to find one example of each of the six market types listed in Table 1 in or near the town or city in which your own school is located.

2. Total revenue is maximized at the point where marginal revenue equals zero. A friend of yours is convinced that this must, therefore, be the profit-maximizing point. Prove to your friend why the profit-maximizing rules cannot hold at the point where total revenue is maximized. Explain to your friend the difference between maximum revenue and maximum profits.

3. Explain what information, if any, each of the following statements gives about the degree of market power the firm may possess.
   a. A firm is the sole supplier of a newly developed product and has 15 years left on a 17-year patent.
   b. A firm has shown a rate of return significantly higher than the average for manufacturing for a five-year period of time.
   c. A firm exists in a market with no entry barriers.
   d. There are no good substitutes for the product produced by a particular industry, so that the market demand curve is very inelastic.
   e. A firm's demand curve lies well below the market demand curve.
   f. There are 20 firms in the market.

4. A friend of yours feels that monopoly power always causes harm. Another friend disagrees, claiming that monopoly power and profits are usually fair rewards for superior performance. To help clarify the issues, draw up a list of ways in which a firm can acquire monopoly power. Briefly explain how each item on your list can help a firm gain monopoly power. Determine whether each specific path to monopoly will involve any social benefits.

5. One problem of monopoly is that consumers are prevented from receiving as much of a particular good as they want.
   a. What is the signal that output is restricted? Explain.
   b. Why is output lower under monopoly than under competition?

6. Suppose that the following statements appear in students' answers to exam questions. Would you give credit for the statements? Why or why not? If not, write a brief explanation for the student showing what was wrong.
   a. X-inefficiency refers to a firm that is too small to achieve economies of scale.
   b. Monopoly will probably show a better performance in terms of the rate of innovation than in the rate of invention.

c. Monopoly power will usually result in a redistribution of income from consumers to the monopolist.
d. Misallocation resulting from monopoly power will be larger when demand is relatively more inelastic.

7. Price discrimination is a common form of behavior for firms with monopoly power.

a. Name three instances of price discrimination that you have personally experienced.
b. Explain the characteristics used to group customers according to elasticities. Explain how these characteristics would cause different demand elasticities.

11

Degrees of Competition

As you read and study this chapter, you will learn:

- the characteristics of dominant firms and their effects
- how oligopoly interdependence may affect pricing behavior
- how monopolistic competition may cause slight deviations from competitive outcomes

Between pure competition and pure monopoly lies the domain of partial competition. It includes the great mass of industrial activity in the modern economy. Whether tiny or huge, simple or complex, most industries are partially competitive rather than at one extreme or the other.

You can readily appreciate from your knowledge of economies of scale that technology and demand shape the structure of firms in a particular industry, and this structure, in turn, may influence both the behavior of individual firms and the economic performance of the industry as a whole. In this chapter, we make the underlying logic more explicit and apply it to cases between pure competition and pure monopoly. In this middle range, the conditions are mixed and the lessons about them are often debatable. Such difficulties make the field a lively one, full of disputes.

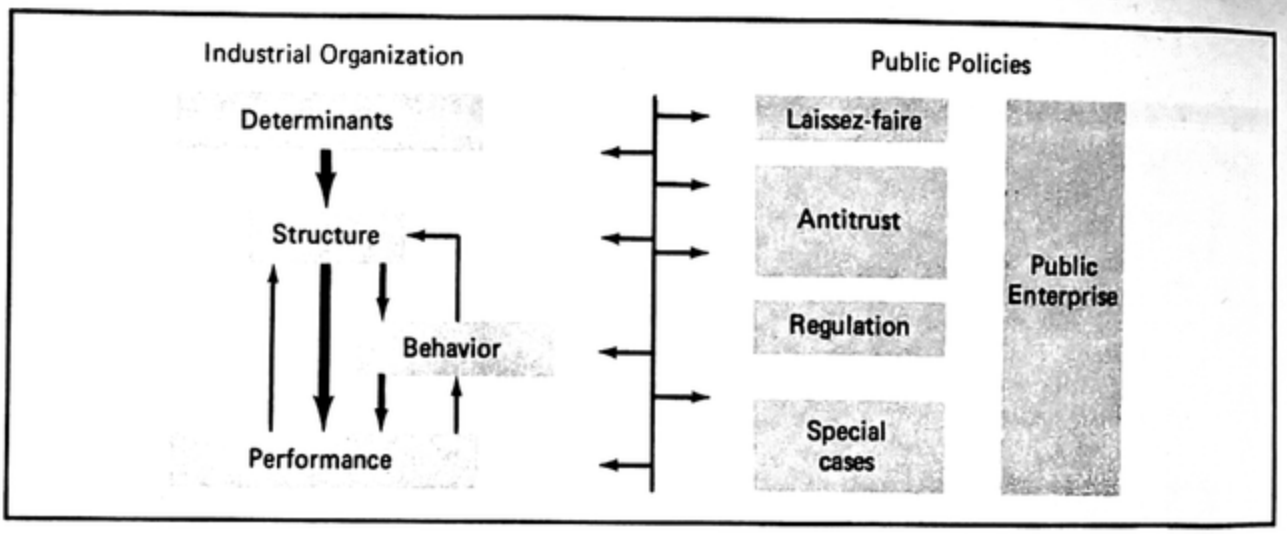


Figure 1 The underlying logic of industrial organization and public policy

The typical industry is on the left-hand side. Causation runs mainly downward, as shown by the thick arrows, from determinants to structure, behavior, and performance. On the right-hand side are the public policies (covered in Chapter 12) that may be taken toward industries with monopoly power. The arrows in the middle go in both directions because policies not only apply to industries but are also affected by them.

Before we discuss this middle range, we should note that this chapter is your first step into a specialized field of economics. Chapters 1–10 were not specialized; they were concerned with the basic theory and facts of economics, which all economists share. But economics has some ten main "applied" fields that deal with specific topics.

"Industrial organization and public policy," which includes the detailed study of markets and policies, is one of these fields. The basic concepts of the industrial organization field are shown concisely in Figure 1. On the left side, each industry is seen as having (1) its own basic underlying conditions, which may shape (2) the industry's structure (its degree of concentration and its barriers against new competition). That structure, in turn, influences (3) the behavior of firms in the industry, and their behavior finally affects (4) how well the industry performs. To show how these concepts are applied to practical cases, we give two contrasting illustrations in Table 1. The fast-food industry is highly competitive, while the automobile industry has been dominated by a few large firms.

Economists have come to divide the degrees of competition into three main categories: dominant firms, oligopoly (from the Greek word for "several sellers"), and monopolistic competition. They were presented in Table 1 of Chapter 10. Although the three categories shade into one another at their edges, each of them has its distinctive concepts. For example, ***dominant firms*** commonly take unilateral actions toward their little competitors. ***Oligopoly*** involves several leading firms instead of one; and these several firms interact in complex ways. **Monopolistic competition** is different from them both, for it involves firms that have only a small degree of monopoly power and little chance to earn excess profits. We will examine each of these categories in this chapter.

The dominant firm

Definition of dominance

Market share A firm is said to be dominant when it has over half of the sales in the market and is more than twice the size of the next largest firm. Note that the dominance is defined primarily by market share. The higher the dominant firm's market share, the closer it comes to being a pure monopoly. To that extent, the firm's demand curve slopes up, perhaps steeply.

Table 1 ***Two case studies of partial competition***

	Fast-Food Restaurants	Automobile Industry
Determinants	The technology gives limited economies of scale, so that minimum efficient scale is reached by the typical local fast-food restaurant.	Economies of scale are significant, requiring plants to have at least an 80,000-car-per-year capacity. Firms need to have at least two lines of car models.
Structure	Low concentration and easy entry.	Medium to high concentration; the largest four firms have about 75 percent of the U.S. market. High entry barriers, except against imports.
Behavior	Flexible pricing; little price discrimination.	Oligopoly pricing, with close interaction among U.S. firms. Mutual reliance on frequent model changes.
Performance	Little excess profit; efficient operations; rapid innovation.	Substantial excess profits (before import pressure during 1978–1981). Inferior arrangements for quality control and for worker incentives. Narrow innovation and neglect of small fuel-efficient designs before 1976.

FOR WO
LIBRARY
Acc No 15358
Date
NAGAR

At high market shares, the dominant firm's demand curve approaches the slope and position of the entire market demand curve.

That slope gives the dominant firm a distinct marginal revenue curve, which lies below its demand curve. The dominant firm therefore acts like a pure monopoly, even though its power over the market is less than complete. There is some competition from the small competitors, but it is usually not severe. Mainly, the dominant firm just sets its profit-maximizing decisions unilaterally, given the degree of monopoly that its demand curve provides.

Entry barriers An entry barrier is any condition that makes it difficult for a firm to enter a market for the first time, so as to become a new competitor. Such entry barriers can reinforce the market power that the dominant firm derives from its market share. There are several types of causes of entry barriers. One is product differentiation, which arises from heavy advertising and from trademarks of brand names. Thus, new entry into the beer, detergent, and toiletries markets is difficult because advertising has wedded many consumers to familiar brands such as Budweiser, Tide, and Right Guard.

Barriers are also caused by large economies of scale, which can force a new entrant to raise large amounts of expensive capital to come in at "minimum efficient scale" where average costs are lowest. Barriers can also come from other causes, such as the need for crucial ores or special skilled workers, which the dominant firm has and new entrants can't get.

Instances and effects of dominance Dominant firms are unusual because a high market share is hard to capture and maintain. Yet, the firms that do get market dominance often become household names. Notice the familiar company

names in Table 2. The computer, car, razor blade, film, or detergent that you buy, the soup you sip, and certain other widely used goods are likely to have been made by dominant firms. Their names and brands are well known precisely because the firms are dominant, producing a large share of the goods in their markets. Many local markets also contain dominant firms. Your local newspaper is probably one, and so perhaps is the biggest local bank, lumberyard, taxi company, and hospital. Judging carefully, you may be able to discover several others, especially if your city is not large.

Dominant firms usually have two effects on prices, similar to those of pure monopoly (recall Chapter 10): (1) They raise the level of their prices, often (though not always) gaining excess profits; and (2) they engage in price discrimination. These traits are normally weaker than would occur under pure monopoly, for dominance is a diluted form of monopoly. The remaining firms do provide a degree of competition. Yet, the patterns are usually similar to what pure monopoly causes.

EXCESS PROFITS arise approximately in the pattern shown in Figure 2. Profitability is commonly measured by the rate of return on equity capital. Market shares are given on the horizontal axis. The statistical pattern has emerged in repeated testing; though there are exceptions, higher market shares correlate closely with higher rates of return. Dominance usually yields progressively higher excess profits, above the cost of capital.

PRICE DISCRIMINATION is also common. Controlling half or more of the sales in the market, dominant firms can often segment the market and set varying price-cost ratios for distinct customer groups. The discrimination will be weaker than it would be under pure monopoly, but the patterns will be similar. For example, General Motors was recently making a profit of over $3,000 per car on its Cadillacs, but only $100 on its small Chevettes. Demand elasticity is low on Cadillacs because buyers will often pay extra for the status they give. But for Chevettes (which face stiff competition from domestic and imported subcompacts), demand is highly elastic. To take another example, IBM has set much higher price-cost margins on its small machines, where competition has been *weakest*, than on its largest computers. And Campbell Soup is said to price its standard tomato soup close to cost, while setting higher price-cost margins on its clam chowder and other fancy soups. These few examples merely illustrate the same result.

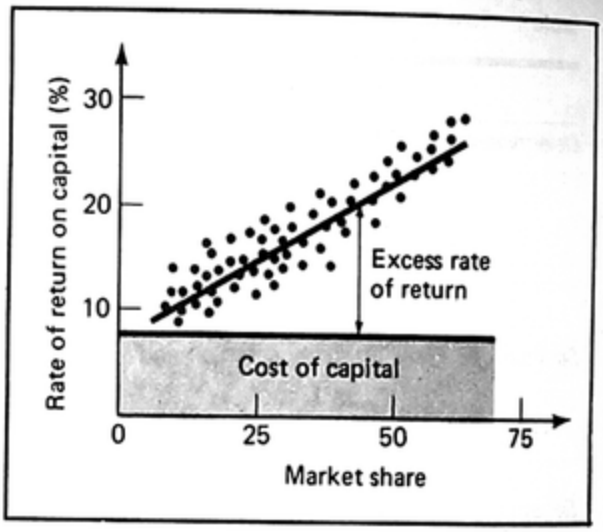


Figure 2 Market shares correlate with rates of profit

Economic theory predicts that companies' market shares are correlated—though not perfectly—with their rates of return. This has been affirmed in statistical testing of large U.S. corporations and other groups of firms. Each dot is for one large company, showing its average market share and rate of return during 1960–1969. There is some variation around the main pattern, caused by a variety of other influences.

Source: W. G. Shepherd, *The Economics of Industrial Organization* (Englewood Cliffs, N.J.: Prentice-Hall, 1979).

Possible causes of dominance

Economies of scale Dominance is often said to reflect the economies of scale arising from modern technology. The dominant firms themselves invariably argue that there is only room for one or two effi-

Table 2 ***A selection of leading instances of dominant firms, oligopolies, and monopolistic competition***

1. Dominant Firms	Markets	The Firm's Average Market Share (%)	Entry Barriers
IBM Corp.	Computers, electric typewriters	60	High
Western Electric Corp.	Telecommunications equipment	95	High
Eastman Kodak Co.	Photographic supplies	60	Medium
Procter & Gamble Co.	Detergents, toiletries	50	Medium
Boeing Corp.	Aircraft	55	High
United Industries	Aircraft engines	50	Medium
Campbell Soup Co.	Canned soups	85	Medium
Gillette Corp.	Razors, toiletries	60	Medium
Wall Street Journal	Business newspapers	65	High
Washington Post	Washington, D.C., area newspapers	84	High

2. Oligopolies	Sales Revenue, 1977 ($ million)	4-Firm Concentration Ratio in Relevant Markets (%)*
Artificial fibers	1,003	80
Automobiles	76,518	84
Flat glass	1,577	90
Batteries	666	87
Glass bottles	3,664	54
Cereal breakfast foods	2,497	89
Newspapers	13,055	90+
Chewing gum	567	93
Cigarettes	6,377	95
Steel	15,331	55
Oil refining	91,688	55
Bearings	2,567	56
Beer	6,652	64
Cement	3,042	56
Fabric weaving	6,325	42

3. Monopolistic Competition	Sales Revenue, 1977 ($ million)	4-Firm Concentration Ratio in Relevant Markets (%)
Movie theaters	2,606	30
Poultry	5,746	16
Yarns	3,846	19
Commercial printing	9,359	18
Knit fabrics	3,169	20
Sheet metalwork	4,863	10
Costume jewelry	816	23
Retail shops	723,134	6
Restaurants	28,470	24
Wood millwork	3,928	14
Dresses	4,188	8

Source: W. G. Shepherd, *The Economics of Industrial Organization* (Englewood Cliffs, N.J.: Prentice-Hall, 1979), Chapter 10, updated using various sources. Adapted by permission of Prentice-Hall, Inc., Englewood Cliffs, N.J.

*The 4-Firm concentration ratio is the share of the market's sales made by the largest 4 firms in the industry.

cient firms in the industry. Therefore, the cost curve shapes you learned in Chapter 8 now become crucial. If scale economies are actually small in those industries (as in Panel I of Figure 3), then the market is *naturally competitive* because there is room for many efficient firms. In that case, any dominance would merely reflect sheer market power. But if economies are larger (as in Panel II), the dominance might be caused by—and therefore provide for—lower average costs.

The issue is clear and important, but the facts are a matter of intense debate. Measuring cost curves is difficult. Technology is usually not simple and rigid, so that costs can't be read off easily from simple formulas. The best research was summarized in Chapter 8. The key question is: Are the dominant firms larger than the optimal scale ("minimum efficient scale")? It turns out that most dominant firms are indeed larger than scale economies seem to require. In some small local markets, such as newspapers, scale economies may indeed be large enough to require high market shares. The markets are small because they are local. Even a modest amount of scale economies will therefore make the firm bulk large in the market, like a medium-sized frog in a small puddle. Yet, the research has not yet reached definitive answers, and conditions can change. As research continues, this consensus—that scale economies do not cause or justify the actual degree of concentration in many national industries—may, of course, be revised.

Yet, there is a contrary view: Modern business has to be big to be efficient. For many decades, some business observers have said that large factories and whole firms are the natural form for mass production and national sales. That may have been true during 1890–1940, at least in some industries, but countertrends toward a smaller optimal scale have quietly set in since the 1940s in many industries.

There are three main reasons for this reversal:

1. The *industrial mix* has changed. There has been a broad shift from "heavy," relatively crude industrial processes to

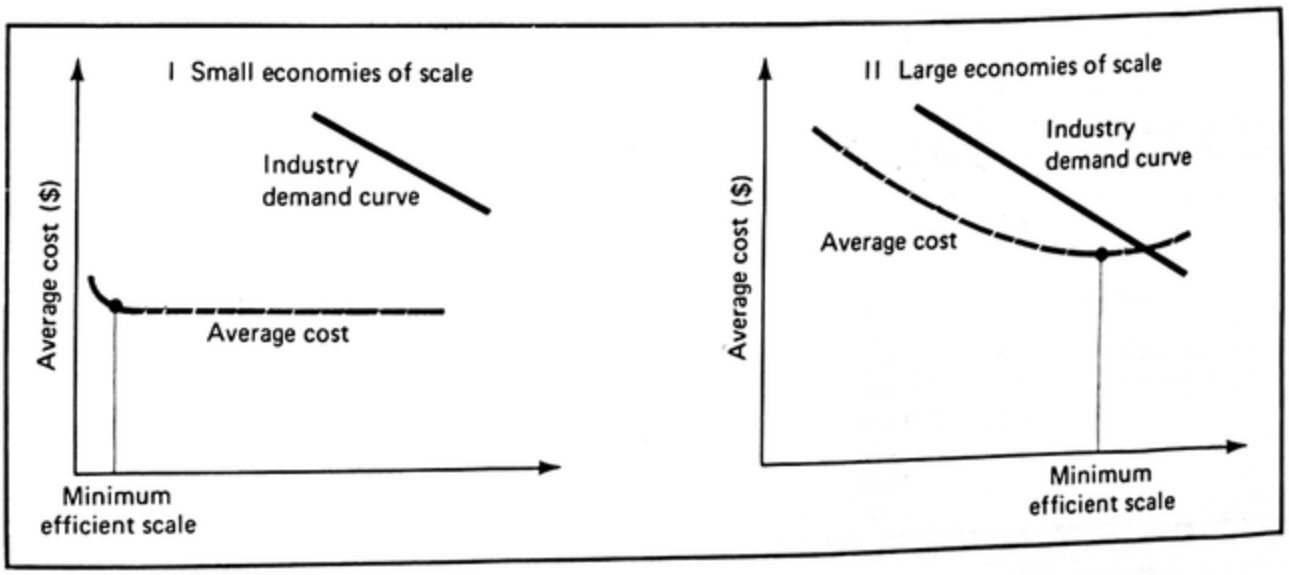


Figure 3 Economies of scale vary between large and small

In Panel I, small economies of scale cause minimum efficient scale (MES) to be only 3 percent of the market. As many as 33 companies can coexist efficiently in such a naturally competitive market. In contrast, Panel II shows an instance where MES is 90 percent of the market. A dominant firm is sure to exist in that case.

more complex, delicate products involving high technology. From such typical products as cast-iron bridges, bricks, and battleships around 1900–1910, the economy's array of products has shifted toward computers, antibiotics, missiles, and stereo sets. The more refined products often require close controls and small-scale assembly, rather than the more crude mass production that could turn out simple industrial products. Therefore, the newer goods are often best produced in smaller factories and firms.

2. *Technology* has shifted broadly in favor of smaller-scale power sources and controls. The power to drive production comes nowadays not from large steam engines but from small electric motors, applied precisely. Small, powerful computers permit exact controls on production, even in small plants. Trucks traveling on the highway network can link numberless small factories much more efficiently than the earlier system of railroads and waterways. In earlier decades, the transport system typically involved a few big factories along the major railroad lines. That favored concentration, rather than a dispersal among many small plants and companies. The telephone system, too, has fostered small-scale efficiency by making close coordination possible even among many separate small firms.

3. *Workers' attitudes* are now more independent, less easy to regiment on a large scale than around 1880–1920, when immigrants were flooding into the labor force. Now, under the pressure of import competition, there is a trend toward more worker responsibility and more flexible work arrangements. These conditions also favor smaller scale.

In light of these trends, we can consider the origins of those firms that have held dominant positions since 1900. Some of them can still be considered dominant in their markets, though many have faded. At least several main causes helped create these firms.

Mergers When two or more competitors combine to form one firm, that is a *horizontal merger* (called "horizontal" because the firms were previously side-by-side competitors). In the three great waves of mergers that we noted in Chapter 7, dominant firms were created, especially in the first wave during 1890–1901. Hundreds of large firms were created, many with 80 or 90 percent of their markets. Usually, these mergers only provided market power, not any true economy of scale.

Many of these instant dominant firms dwindled or collapsed under the pressure of new competition or owing to their own incompetence. Yet, others persisted, some of them even down to the present. Prominent cases of such merger-created dominant firms include U.S. Steel (1901–1910), American Tobacco (1900–1915), United Shoe Machinery (1890–1960), and General Electric (1900–1930).

Patents and trademarks Patents are official monopolies issued by the government to inventors. They protect new inventions during a 17-year period, so that the inventor can gain a high reward in selling the new product. The economics of patents are discussed in Chapter 16. For now, note that many dominant firms originated with one or more crucial patents. Major instances (with their periods of dominance) are the Aluminum Company of America (1895–1950), Gillette Company (1903 to the present), Kellogg's in corn flakes (1890–1930s), Xerox Corporation in copiers (1961–1975), Polaroid Corporation in

Newspapers: Dominance and Scale Economies

The true markets for most newspapers are limited to their cities and surrounding areas. Within these markets, it is common for one firm to have high mar-

Dominant firms in selected newspaper markets 1980–1981

Metropolitan Areas	Circulation of Copies, 1980–1981 (12-month average)	Newspapers' Share of Market (%)
Des Moines, Iowa	107,095	
*Register, Tribune**	106,803	99.7
New Orleans, Louisiana	245,480	
*Times-Picayune, States-Item**	238,230	97.0
Miami, Florida	331,843	
*Herald, News**	312,765	94.3
Milwaukee, Wisconsin	442,151	
*Journal, Sentinel**	398,491	90.1
Kansas City, Kansas-Missouri	562,000	
*Star, Times**	503,910	89.7
Atlanta, Georgia	385,513	
*Journal, Constitution**	341,562	88.6
Washington, D.C.	616,493	
Post	517,989	84.0
St. Louis, Missouri	519,718	
*Post-Dispatch, Globe-Democrat**	435,405	83.8
Phoenix, Arizona	391,688	
*Republic, Gazette**	317,511	81.1
Austin, Texas	139,201	
American-Statesman	111,113	79.8
Akron, Ohio	192,175	
Beacon Journal	139,978	72.8
Pittsburgh, Pennsylvania	658,552	
*Press, Post-Gazette**	411,207	62.4
Baltimore, Maryland	548,790	
Sun	328,417	59.8
News American	137,645	25.1
Los Angeles, California	1,516,102	
Times	690,749	45.6
Herald Examiner	119,245	7.9
Chicago, Illinois	1,596,593	
Tribune	657,951	41.2
Sun-Times	637,381	39.9
Boston, Massachusetts	1,095,072	
Globe	410,015	37.4
Herald American	191,202	17.5

Source: American Newspaper Markets, Inc., *Circulation '81/'82* (Malibu, Cal., 1981).

*Newspapers whose names are separated by commas are owned by one firm (e.g., the Des Moines *Register* and *Tribune*).

ket shares. The table presents a selection of such dominant firms, plus a few with lower market shares.

These are familiar papers, some of them civic-minded and long established. They vary in tone, political slant, and quality; most of them are highly profitable. In any event, they are all dominant firms holding a high degree of market power. Since many of their owners also own local television and radio stations, their market power in the entire local media market is often very great.

These newspapers are survivors from earlier eras in which most large cities had three, four, or more newspapers. What has caused the decline in numbers of newspapers? The main causes are rising economies of scale in printing and delivery, and the tendency for advertisers to cluster their spending in only one paper. Even the largest "second" newspapers have been squeezed, such as the venerable *Washington Star*, which folded in 1981.

instant photography (1948–1980), and many medical drugs since the 1940s.

Monopolizing tactics, including price discrimination By sheer tenacity and various tactics, a number of commercial geniuses have built up dominant firms. As we saw in the last chapter, John D. Rockefeller and Standard Oil is a leading instance, involving price discrimination. Others include Eastman Kodak in photographic film (1895 to the present), General Motors in automobiles (1930–1950), IBM in computers (1955 to the present), and United Fruit in bananas (1890–1960).

Scale economies Economies of scale have sometimes been important at the beginning of an industry, when one firm got a firm foothold. It is less usual for scale economies to preserve a dominant position in a mature industry. Among the few present-day examples are newspapers in many cities. Yet, such cases are often debatable, for often there are small competitors whose existence suggests that the scale economies are not conclusive.

Oligopoly

Concentration and leading firms

The economist's basic criterion in defining oligopoly is ***concentration,*** and the usual practical measure of it is the *concentration ratio.* That ratio is the combined share of the market's sales that is held by the largest four firms in the market.* For example, if market shares in the United States automobile market are General Motors 45 percent, Ford 20 percent, Chrysler 10 percent, and Toyota 4 percent, then the four-firm concentration ratio is 79 percent.

Concentration obviously can vary continuously from near zero for a purely competitive market all the way up to 100 percent when there are only four firms or fewer. One therefore speaks of concentration in numerical terms, such as 35, or 62, or 93 percent. Table 2 shows some of this

*In the United States, concentration is reported only for the top 4, 8, 12, and 20 firms because the Census Bureau is forbidden to disclose facts that could reveal individual companies' conditions. Three-firm concentration ratios would be valuable to have, but—it is claimed—two of the three firms might get together, share their data, and learn the market share of the third! So no three-firm ratios are published, and scholars must make do with four-firm ratios.

variety. But economists divide this continuous range into several approximate categories, in order to clarify certain distinct features. They usually apply the phrase ***tight oligopoly*** to markets whose four-firm concentration is above 60 percent or so. By contrast, ***loose oligopoly*** occurs when concentration is in the range of about 20 to 40 percent. When concentration is below 20 percent, economists usually define the market condition as monopolistic competition. That is treated in the next section.

The typical oligopoly has several leading firms plus a fringe of little competitors. Those leading firms are rarely equal in size, as in Panel I of Figure 4. Instead, they normally taper down from the biggest one, as in Panel II of Figure 4, with each firm substantially smaller than the next. It is often not exactly clear where the oligopoly firms end and the fringe firms begin. Nonetheless, the key economic feature of oligopoly is the presence of a group of leading firms. This leads to two distinctive conditions.

Interdependence One of the distinctive conditions about the oligopoly group is the ***interdependence*** among its members. They must be constantly aware of one another's actions, planning their own moves carefully and ready to react to one another's tactics. Therefore, oligopoly is permeated with strategy. Actions cannot be simple unilateral steps, as they are in pure competition, pure monopoly, or dominance. These strategic conditions will be presented later in this chapter.

Tight oligopoly is often like a chess game or a war. Each firm's choices hinge on what it expects the others to do, in response to its actions or in following their own strategies. Like a chess player or a general plotting a military campaign, an oligopolist often needs to think three or four moves ahead—in short, to have a strategy. The oligopolist tries to anticipate its rivals' reaction to its own actions. Its own choices, therefore, often depend crucially on what its rivals' policies are likely to be. Moreover, each of the other firms

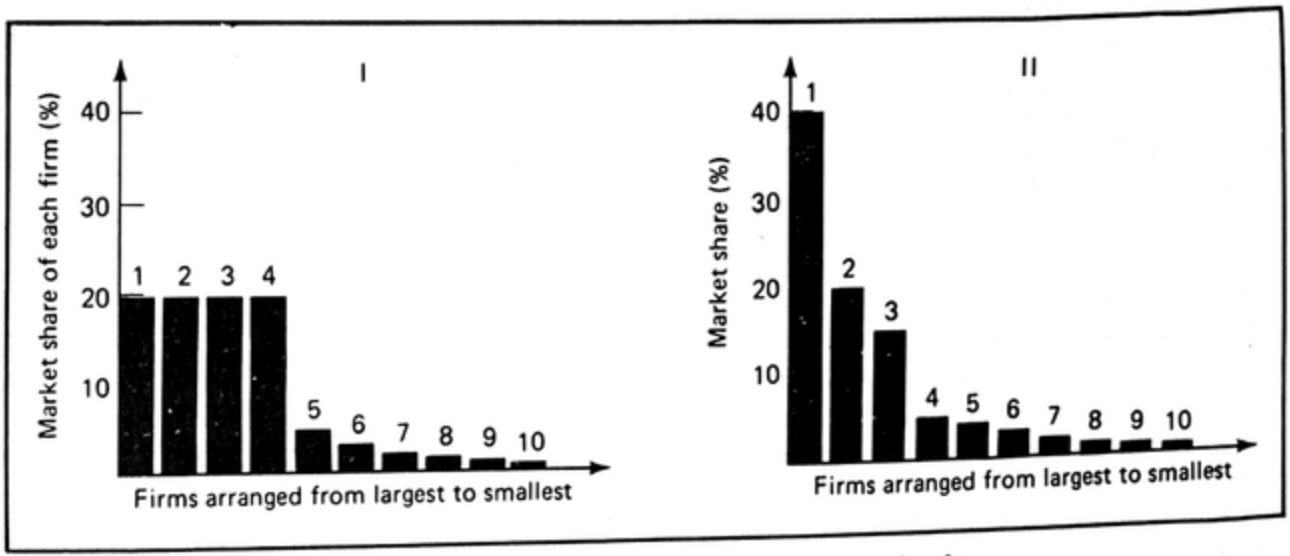


Figure 4 Oligopolists may be of similar sizes, but they usually vary in size
Panel I shows a tight oligopoly with four equal-sized firms. In Panel II, by contrast, the oligopolists have sharply varying sizes, even though their total concentration, 80 percent, is just the same as in Panel I.
The tapered pattern in Panel II is far more common in actual oligopolies than is the equal-size group in Panel I.

The Schumpeterian Process

The strongest argument that big business is highly competitive has come from one of the most colorful and erudite of economists, Joseph A. Schumpeter (1883–1949). An Austrian, he was a young diplomat with a stable of horses in Egypt, a politician, and then, after fleeing Hitler in 1933, the reigning economic theorist at Harvard for 16 years. Deeply conservative, he sought to defend capitalism, even its monopolies. His striking image of "creative destruction"—published in 1944—was and is a dramatic dissent from the prevailing view of neo-classical economists.*



Competition and progress go together, he said, but in a series of temporary monopolies. The Schumpeterian version of competition is almost the exact reverse, point by point, of the neoclassical equilibrium analysis. At each period of time, each market might be dominated by one firm, which earns monopoly profits. But these high profits attract other large firms, one of which will innovate and displace the first dominant firm. The new dominant firm then thrives, but it too is pushed aside.

This cycle of "creative destruction" continues: innovation, dominance, monopoly profits, new innovation, a new dominant firm, and on and on. As time passes, the average degree of monopoly profits might be small—surely smaller than the innovators hoped. Meanwhile, the cycle might generate benefits of technical progress far exceeding any costs of marginal misallocation which are caused as market power comes and goes.

Capitalism, Socialism and Democracy (New York: Harper & Row, 1944).

must think strategically too, so that the whole process is complex and thoroughly problematic.

Oligopoly thus differs from all other market forms. Under monopoly, dominance, pure competition, and (as we will soon show) monopolistic competition, each firm merely finds and achieves its profit-maximizing levels of price and output. It ignores any possible responses by specific other firms.

Variety Oligopoly's other distinctive trait is its variety. It embraces a remarkable diversity of conditions. *Concentration* may range from 20 percent on up. The leading

firms may be about *equal* to one another in size or, instead, even more strongly *different* than in Panel II of Figure 4. The *products* may be "homogeneous" (like cement and lumber) or "differentiated" (like electrical generating plants and brands of cereal or beer). *Entry barriers* to the market may be high, medium, or low.

Oligopoly's main lines Since oligopoly embraces such a wide variety, there can be no single economic "model" for it. Economists have only been able to develop several general lessons and methods for showing oligopoly's main lines. These concern:

1. The oligopolists' conflicting incentives, either to compete or collude with one another.
2. The contrast of outcomes between tight and loose oligopoly.
3. The central tendency that usually emerges under oligopoly.
4. Ways to explain why oligopoly prices are often rigid over long periods of time.
5. Whether economies of scale have been the main cause of oligopoly concentration.

Conflicting incentives can make oligopoly unstable

Each firm in an oligopoly has mixed incentives toward its several rivals. Depending on how the balance tips, a firm may fight, or cooperate, or reach some mix of actions toward its fellow oligopolists.

Competing Each firm could compete intensely against its rivals, seeking every way to defeat them and maximize its own profits. Of course, an aggressive firm must expect sharp retaliation from the others. Its own hostile actions may force the others to respond in kind, even if they were not equally hostile and aggressive in the beginning.

Colluding Yet, collusion is also attractive. Each oligopolist knows that if all the firms in the industry cooperate, they can maximize their total profits. At the extreme, they might make as much total profit as if they were united in a single complete monopoly. Such "joint profits" will greatly exceed the sum of the individual profits that firms make when they are fighting one another. For example, three beer companies agree to hold prices at $2.70 per six-pack, well above the average cost of $2.30 per six-pack. Sales are 37.5 million six-packs, which yields them excess profits of $50 million each, or $150 million in total. If they had competed firecely, the price might have been driven down to $2.45 per six-pack, raising the sales to 60 million six-packs. The result would provide them only $3 million apiece in excess profits, for a total of $9 million.

As with these firms, so with the common run of oligopolists: The incentive is to cooperate, not compete. The rewards from cooperation and collusion depend largely on how concentrated the market is.

The higher the concentration, the stronger are the firms' incentives and opportunities to cooperate successfully and thus to maximize their joint profits. If there are only two or three firms, this urge to collude is strong and often compelling. Yet, even when concentration is low, the incentive to collude is still present. Oligopolists can always gain by setting up a price-fixing ring and raising the price—*if* they can prevent price cutting.

But often they cannot suppress the urge to compete. Even if the oligopolists do raise price so as to achieve maximum joint profits, each firm still has the contrary incentive to compete. Once the high collusive price is set, each firm could gain

by secretly cutting its own price just below the jointly agreed upon price. The slightly lower price will take away large amounts of sales from its rivals, and so this one firm's profits will increase. As soon as other firms discover why they are losing sales, they will be under pressure to cut their own prices or at least to try to penalize the "chiseler." Cooperation may well collapse and a price war may break out.

Accordingly, oligopolists' collusion is often unstable, breaking apart from its own inner tensions. Thus, in the example of the three beer firms, one of them could gain $10 million in profits by cutting price to $2.65 per six-pack *if* the other two stay at the $2.70 price. For instance, the price cutter could raise its sales to 20 million six-packs; at a profit of 35 cents per six-pack, the firm's excess profits would now be $70 million.

All price-fixing firms share this temptation to cut the price. But they also know that the other two will probably retaliate. The situation mingles rewards for both price cutting and price fixing.

Oligopolistic industries often veer between being restrictive and stagnant when cooperation holds, and being aggressive and progressive when competition breaks out. Oligopoly frequently is like Dr. Jekyll and Mr. Hyde. Or, rather, it is like a person who, constantly torn between the temptation to sin and belief in a moral code, resists at some times but gives in at others. There is always the tension between opposing choices.

Each oligopolistic setting provides several possible outcomes. The specific results of each one are usually *indeterminant*—that is, they cannot be predicted in advance. The structures are too diverse, and the attitudes of the firms' managers are varied and unpredictable. Thus, the outcomes vary. In one oligopolistic industry, the firms will settle into snug cooperation and act like a monopoly with high prices and little innovation. In another oligopoly, the firms will engage in endless warfare, with low prices and frantic innovation. A third oligopolistic industry will alternate between the extremes.

Despite this variation of specific cases, there are some predictable basic patterns that favor collusion. When the following three conditions occur, collusion is likely to stick.

1. Similarity of the firms' conditions If the firms have similar demand conditions and/or similar cost conditions, they will be more able and more likely to cooperate. Their interests will coincide, and they can have more confidence that cooperation will last. That very confidence will deepen their mutual trust and make collusion among them more likely to succeed.

For example, if there are three oligopolists in the copper industry with identical marginal costs of 68 cents per pound, then they may easily agree on 97 cents per pound as the best price. But if their costs differ—at 51 cents, 63 cents, and 82 cents—they will have trouble setting one price. At a 97-cent price, the lowest-cost firm (with its 51-cent marginal cost) would wish to cut price perhaps to 68 cents, while the highest-cost firm (with marginal cost at 82 cents) would fight such a move. In short, differences breed discord, whereas similarity breeds cooperation.

2. Familiarity over time Each firm's managers get to know the other companies as time passes, and they learn to judge and predict one another's behavior more accurately. Misunderstandings become less likely, and mutual trust grows. The oligopolies in older industries tend to have a clubbier, more comfortable atmosphere. When new managers take over in one firm or another, things may be less stable for a

while. But further experience tends to restore mutual understanding.

3. Concentration. The likelihood of cooperation varies closely with the *degree of concentration.* Higher concentration breeds more collusion, for two main reasons. First, collusion is easier when there are fewer firms controlling the bulk of the market. The few firms can organize, understand, and enforce their mutual agreements more thoroughly. Moreover, the leaders—having most of the market—face little pressure from those small fringe firms that are outside the price ring. Those fringe firms' price actions can have only a mild effect on the leading firms' market share. Second, price cutting by any renegade oligopolist is easier to discover and penalize when there are only a few firms. If there are only three firms, the other two will quickly know that the third firm is the chiseler. In contrast, if there are 15 or 20 firms involved, any one of them or several—or all—will be more sorely tempted to chisel, since each one can expect to succeed for a longer time before being discovered.

Types of collusion

The kinds of collusion that may occur in oligopolies range from tight, explicit collusion to informal, loose arrangements.

Explicit, formal collusion If price fixing is legal, the price fixing in tight oligopolies can be so complete that it approaches the level that a pure monopoly would achieve. *Cartels* may be formed. A cartel is a formal organization created by companies to manage their cooperation. It fixes prices and enforces penalties against members who violate the agreement. The cartel may also set output quotas, control investments, and pool profits. Most cartels have existed in Western European countries and certain international markets, such as OPEC in the world oil market.

Price fixing has been against the law since about 1900 in most U.S. industries, under Section 1 of the Sherman Act and various state antitrust laws (see Chapter 12 for details about antitrust). The U.S. antitrust laws, therefore, shift the margin of choice away from collusion and toward competition in most U.S. oligopolies. But there is some hidden price fixing, nonetheless, done through secret meetings, phone calls, and other covert ways. Indeed price fixing is a way of life in many industries, as the leading business magazines and newspapers have occasionally noted. Table 3 presents some typical instances, drawn from a wide variety of markets. Each year, scores of antitrust cases turn up many more examples.

Tacit collusion Price fixing can also occur in a milder form, called *tacit collusion,* or parallel pricing, or price signaling. The oligopolistic firms do not conspire directly or sign binding agreements, mainly because it is against the law to do so. But a firm can give indirect hints and signals of its preferred price levels. Then all the other firms simply go along with the same price changes. Often, a cooperative price is reached, just as if formal collusion had occurred.

One version of this tacit collusion is called *price leadership.* In it, one firm periodically judges the best new joint-maximizing price for them all and then sets that price. The others simply follow the leader, quickly matching its price. The pattern then is: long periods of stable prices, punctuated by simultaneous price jumps, usually led by the same firm. Such lockstep, stair-step pricing strongly suggests some degree of tacit collusion. In practice, prices usually do not rise in such a starkly rigid form. Instead, the firms often shift their prices at about the same time, to

Table 3 ***Typical instances of price fixing in recent years***

Product	Geographic Scope	4-Firm Concentration (combined share of the largest 4 firms) (%)	Number of Conspiring Firms
Bedsprings	National	60	10
Self-locking units	National	97	4
Women's swimsuits	National	65	9
Carbon steel sheets	National	69	10
Commercial baking flour	Regional	50	9
Gasoline	Regional	45	12
Liquid asphalt	Regional	56	20
Book matches	National	77	10
Concrete pipe	Regional	100	4
Linen supplies	Local	49	31
Plumbing fixtures	National	76	7
Class rings	Regional	95	3
Tickets	Regional	97	3
Baked goods	Regional	46	7
Industrial chemicals	Local	90	5
Armored-car services	National	99	3
Vending machines	Local	93	6
Ready-mix concrete	Local	86	9
Wrought-steel wheels	National	85	5
Metal library shelving	National	60	7

Source: George A. Hay and Daniel Kelly, "An Empirical Survey of Price Fixing Conspiracies," *Journal of Law and Economics*, 1974, pp. 13–38. © 1974, The University of Chicago Press.

about the same levels. The economist then has to judge whether tacit collusion has really occurred, or if instead the shifts largely reflect changes in costs.

Tight oligopoly crystallizes such indirect cooperation much more fully than does loose oligopoly. When the few rivals control nearly all of the industry, tacit collusion can often be nearly as complete as with a full-blown cartel (or even with pure monopoly). Loose oligopoly, in contrast, has too many "leading" firms, and their collective market share (that is, the concentration ratio) is less than 40 percent. Thus, loose oligopoly is usually a scene of chronic strife, often degenerating into flexible, competitive pricing action.

The contrast is not absolute, however. Tight oligopolies often undergo bouts of fierce competition when collusion collapses, and loose oligopolies can sometimes be effectively collusive. But the general contrast is valid. It marks out tight oligopoly as a special problem, even where price fixing is illegal.

The central tendency under oligopoly

The central tendency in oligopolistic markets is for firms to converge on identical prices and product features. Collusive prices under tight oligopoly is just one part of this fundamental tendency. We now explain this central tendency in broader terms.

The customers in a market are often distributed along a range: of geographical areas, of product types, or of time periods. Typically, the distribution has a central cluster with two tails, as shown in Figure

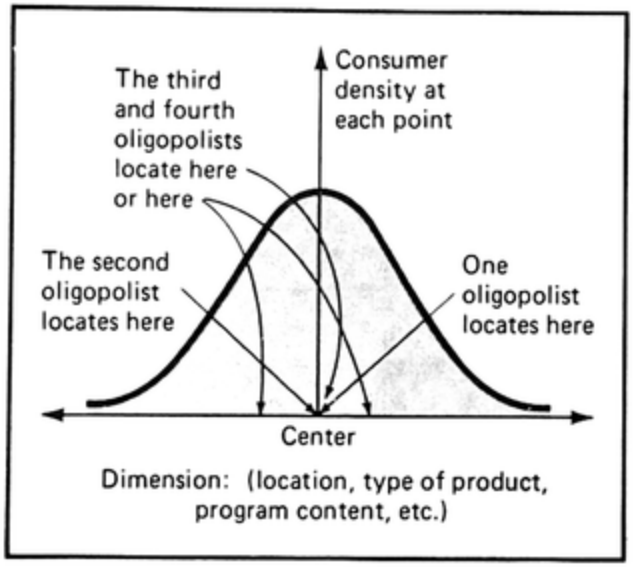


Figure 5 Under tight oligopoly, the firms cluster at the center of the market

Customers are arrayed along a dimension, such as geographical location or type of product. Most of them cluster in the middle. Two oligopolists will choose the exact middle, back to back. A third and fourth firm will also probably crowd into the center. Only as more oligopolists arise will they possibly locate in the other parts of the distribution, catering to the variety of customers.

5. For example, a country town is spread along the main road, with a clustering of population at the center.

If there are just two competitors ("duopolists"), they will locate at the exact center of the distribution. Thus, two gasoline stations, restaurants, bookstores, department stores, or banks will usually be found close to each other at the middle of town. In practical terms, they will both locate at the main intersection downtown. Equally, two television networks will tend to produce the same sort of mass programs, aiming for the center of the distribution of program preferences. The two political parties, Republicans and Democrats, tend to offer similar programs to voters, rather than take radically opposed positions. Therefore, duopoly exhibits a strong tendency for the two oligopolists to adopt uniform positions at the center of the market.

This tendency continues even when there are three or more oligopolists, though it weakens as the oligopolistic group gets bigger. Consider what happens if other firms enter the market. A third oligopolist would also move to the middle of the market, for that is where it can hope to attract the most demand. Additional firms, however, might begin to move away from dead center to seek out special customer groups. For example, a fifth TV network might specialize in the arts, popular sports, or news. That is, indeed, what has happened as specialized cable and satellite networks have emerged since 1978, offering all-day sports and news coverage. Radio stations have also specialized as the old network dominance has been diluted. Every city has several types of popular music stations, each specializing in one "sound" rather than all offering the same music.

Rigid prices: kinked demand curves

Oligopoly tends toward rigid prices, changed relatively infrequently (as noted just above). For example, steel and automobile prices in the 1950s and 1960s followed a distinct stair-step pattern: constant for 12 months, and then raised uniformly by all firms in August of each year. An ingenious analysis to explain this rigidity—which economists call the "kinked" demand curve, because the curve, indeed, has a kink in it—was developed in the 1930s and is still the most widely accepted single model of oligopoly. It is particularly valuable for showing the kind of pricing dilemma that oligopolists face once an industry-wide price has been established.

Its general form is illustrated in Panel I of Figure 6. Panel II gives more details about how it is derived. Its logic can be grasped by putting yourself in the oligopolist's shoes. Suppose that you run an oli-

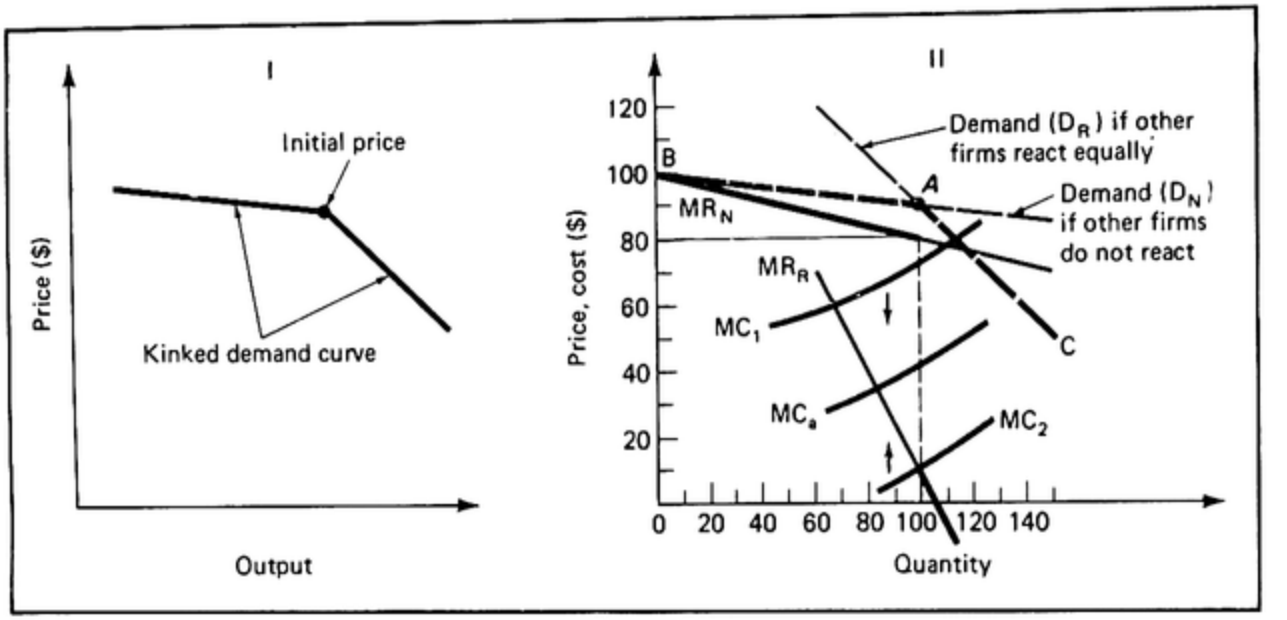


Figure 6 The conventional kinked demand curve for a timid, pessimistic oligopolist

gopolistic firm and are trying to determine the results of changing your price. As in a chess game, your strategy—your choice of a price—depends crucially on the reactions of the other firms in the industry. In fact, you can't determine the demand curve for your firm until you make some assumptions about your rivals' responses. Panel II of Figure 6 illustrates one possible set of such assumptions. Note that the figure doesn't indicate what the rivals will do, only what the firm *expects* them to do.

For example, if your rivals do *not* change price when you do, you will lose a lot of sales to them if you raise the price, and gain a lot of sales if you lower it. In other words, you will face a relatively *elastic* demand curve, such as D_N in Panel II of Figure 6. On the other hand, suppose that your rivals match all of your price changes. Then raising your price won't lure customers away from your rivals. Your demand curve will be relatively *inelastic*, such as D_R in Panel II. Table 4 illustrates these conditions.

Which assumption do you make? Will your rivals match your price changes? Suppose that you take a timid, pessimistic approach. After all, it's better to be pessimistic and pleasantly surprised than to be optimistic and find things turning out worse than you had anticipated.

Therefore, you start at the current price and output level, represented by Point *A* in Panel II of Figure 6, and ask yourself what the worst reaction of your rivals could be to an increase in your price. Clearly, it would be simply to leave their price unchanged. Then, with your own higher price, you would lose a lot of sales to your rivals. The gain from a higher price would be more than outweighed by a loss in sales. In other words, if you raise your price, you perceive that you will be operating on a portion of the elastic demand curve or D_N, such as the *AB* portion: An increase in price results in a substantial drop in quantity.

Now suppose you begin again at Point *A* and consider a price *cut*. Here the worst response of your rivals would be to match your price cut. Now you're stuck with a lower price, but hardly any increase in quantity sold. After all, your lower price won't help you to draw customers from your rivals if they, too, are charging the new lower price. If you lower your price, then, you expect to be operating on a portion of the inelastic demand curve or D_R, such as the *AC* portion: A reduction in

Table 4 ***An illustration of the kinked demand curve***

I. With the Conventional Kink: A Pessimistic Oligopolist

Increasing Price				Cutting Price			
Price	Units Sold	Total Revenue	Marginal Revenue	Price	Units Sold	Total Revenue	Marginal Revenue
$100	0	$ 0		$90	100	$ 9000	
			98				$ 6
98	20	1960		86	105	9030	
			94				−2
96	40	3840		82	110	9020	
			90				−10
94	60	5640		78	115	8970	
			86				−18
92	80	7360		74	120	8880	
			82				−26
90	100	9000		70	125	8750	

II. With the Reverse Kink: An Optimistic Oligopolist

Increasing Price				Cutting Price			
Price	Units Sold	Total Revenue	Marginal Revenue	Price	Units Sold	Total Revenue	Marginal Revenue
$110	75	$8250		$90	100	$ 9000	
			$46				$78
106	80	8480		88	120	10560	
			38				74
102	85	8670		86	140	12040	
			30				72
98	90	8820		84	160	13440	
			22				68
94	95	8930		82	180	14760	
			14				64
90	100	9000		80	200	16000	

price results in only a small increase in quantity sold.

Your pessimistic assumptions about your rivals' reactions result in a *kinked demand curve, BAC* in the diagram. The kink occurs because you expect a relatively elastic demand curve at prices higher than Point *A* and a relatively inelastic demand curve at prices below Point *A*. If you sketch in the marginal revenue curve for this kinked demand curve, you discover another oddity. As you jump from the relatively *elastic* demand curve segment to the left of Point *A* to the relatively *inelastic* demand curve part to the right of Point *A*, you must also jump from the corresponding *elastic* marginal revenue curve segment down to the *inelastic* marginal revenue curve part to the right. To do this, you travel down an actual discontinuity in marginal revenue, shown by the vertical dashed line at output Q_0 from $80 down to $10. (Try drawing this yourself to verify the result.)

With the demand and marginal revenue schedules sketched out, you're ready to apply the $MC = MR$ rule of profit maximization. But now you encounter a serious problem. Suppose that marginal cost lies *within* the area between MC_1 and MC_2, such as MC_a in Panel II. You can't apply the $MC = MR$ rule in the usual way, since there is no marginal revenue curve in the interval between $80 and $10. Because the marginal cost curve MC_a does not intersect the *MR* curve, a policy of no action is best. The producer will remain at Point *A*. Thus, even when changing costs sharply shift the *MC* curve, the oligopolist is not likely to change its price and output levels.

The kink can clarify the long periods of rigid prices that occur in some oligopolies. It also helps one understand why oligopolists often maintain the same pattern of market shares without changes over long periods, as, for example, when the four leading U.S. meatpackers kept constant market shares from 1890 to 1920.

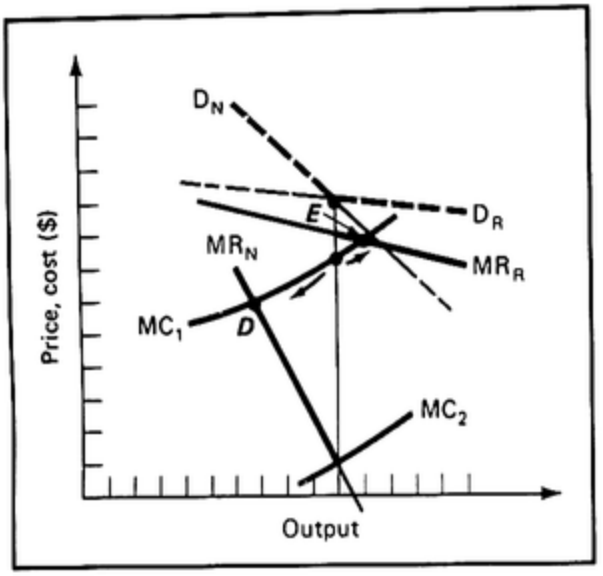


Figure 7 The reverse kinked demand curve for an aggressive, optimistic oligopolist

But kinked demand curves can lead to other conclusions. Suppose the oligopolist has an aggressive and optimistic attitude. It expects the *best* reactions to its price changes, not the worst: Its rivals will match its price rises but not its price cuts. That will cause the *reverse* kink to hold, as shown in Figure 7 (Table 4 gives the numbers for it). For price rises, the firm expects not to lose much quantity. For a price cut, the firm expects to gain a sharply higher quantity.

Now marginal cost equals the expected marginal revenue at both Points *D* and *E* in Figure 7. These points both differ from the original price output levels. Therefore, thinking positively, the firm will now definitely change its price, away from the original level, choosing either Point *D* or *E*. Though we cannot predict which way the optimistic oligopolist will go—whether it will cut or raise its price—the firm *will* change its price and, therefore, also force its rivals to change one way or the other.

Note that these two contrasting results based on pessimism and optimism illustrate the importance of attitudes. An old, familiar group of rivals may have learned to coexist and avoid price changes; each fears to upset the applecart. That situation gives the usual kink, which we explained first. But a brash newcomer, or a change of attitude by one of the firms based on aggressive optimism, can upset the equilibrium. The reverse kink shows that. The resulting sequence of moves among the firms can vary, but a stable equilibrium will recur only when no firm has a reverse kink.

It is important not to take the theory of the kinked demand curve *too* literally. The managers of oligopolistic firms do not spend their time staring at diagrams of discontinuous marginal revenue schedules. But their choices come out *as if* they did this sort of analysis. The theory's value comes from the fact that it illustrates the kind of interdependence that oligopolists face. The effects of future price changes can never be anticipated with certainty, since rivals' reactions are often unpredictable. By illustrating this interdependence and uncertainty, the kinked demand curve is useful for understanding oligopoly.

Economies of scale: a cause of oligopoly?

Oligopolistic concentration is not strongly influenced by economies of scale. The situation differs only in degree from that presented for dominant market shares. Oligopoly includes more cases where market shares of the leading firm are not much above the levels that scale economies make necessary. Thus, many oligopolists in medium and small markets have market shares of 10 or 15 percent, while minimum efficient scale is 8 to 12 percent. "Excess" concentration, therefore, is often not large.

Yet, there are many cases when oligopolists have a larger market share than required for efficiency. Often the biggest oligopolist has a market share of 30 to 40

percent, well above the MES of 10 or 15 percent. Therefore, on the whole, much of the concentration found in oligopoly markets is not required or justified by the economies of scale.

Monopolistic competition

The lower ranges of loose oligopoly shade into another market type, called ***monopolistic competition.*** It has low levels of concentration, but each firm has a slight degree of monopoly. Therefore, economists treat this market structure as a highly diluted form of monopoly, in which firms' demand curves have only a slight downward slope. No firm's market share is more than 10 percent.

The distinctive features of monopolistic competition are as follows:

1. There is some *product differentiation,* which means that consumers can develop preferences among the sellers. This slight degree of market power gives the firm's demand curve a slight downward slope, as illustrated in Figure 8. The product differentiation can occur either (1) because the products themselves *differ physically or in brand images* (like various brands of bread, jewelry, or shirts); or (2) because of the sellers' *locations* (as when a local grocery store, hotel, or restaurant is convenient to a neighborhood).
2. There is *free entry* into the market. New firms enter whenever any excess profit (above the normal competitive rate) is being made in the industry.
3. There is *no interdependence* among individual firms. No firms have large enough market shares to influence the rest of the market. Each firm merely feels the competitive pressure from all of the many other firms in the market.

These conditions are common among retail outlets and in other markets, as shown in the third group of markets in Table 2. A typical case of monopolistic competition is a grocery or clothing store, with

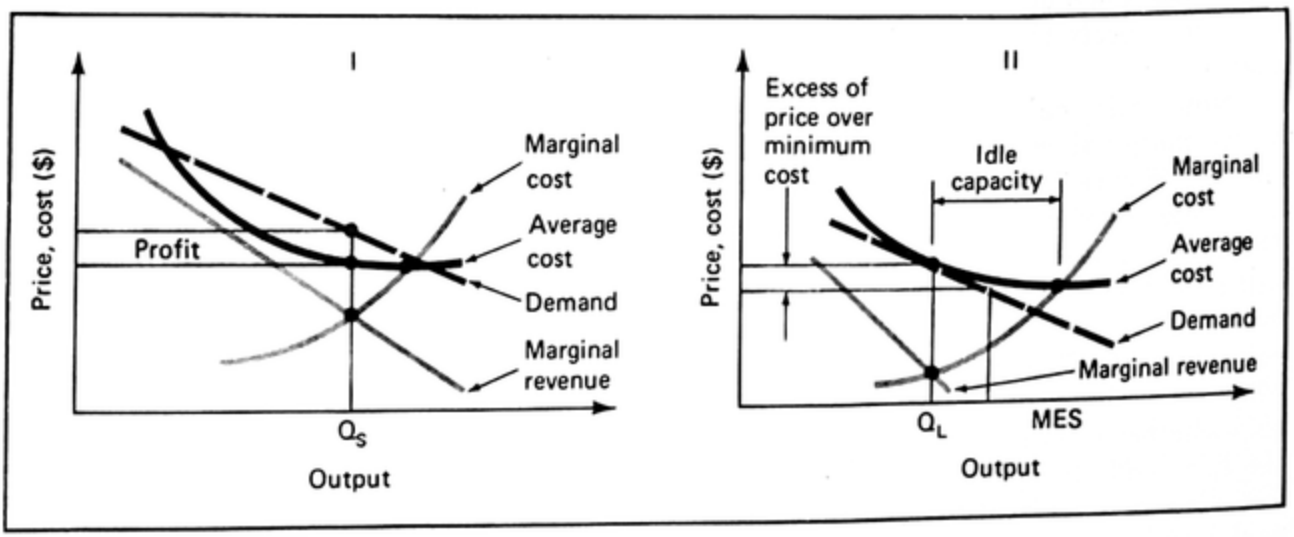


Figure 8 Monopolistic competition
Demand is highly elastic. Profits may occur (Panel I). But soon demand is forced down until the average cost curve just touches the demand curve (Panel II). There is no extra profit, but price is above minimum average cost, and there is idle capacity.

some loyal clientele in its neighborhood but steady competition from many other stores farther away. The firm's demand is highly, but not infinitely, elastic. Because the demand curve is nearly flat, the firm has only a little room for choice.

In the short run, the situation in Panel I of Figure 8 may hold. The demand curve may lie above the average cost curve. That permits the firm to earn short-run excess profits, as shown when it chooses the output Q_S. But then free entry takes its toll. New firms, seeing that excess profits are being made, enter the market, and that forces down this firm's demand curve until it is just tangent to the average cost curve.

In Panel II, the demand curve is nowhere above the cost curve, so that no excess profits are possible. The firm can just survive at output Q_L—where marginal revenue equals marginal cost—barely earning the competitive rate of profit. The power of monopolistic competition eliminates long-run excess profits, and inelastic demand can result in no excess profits.

Yet, monopolistic competition does cause two deviations from the efficient results of pure competition (recall Chapter 9). *First*, cost and price will both be slightly higher than under perfect competition (which settles at MES in Panel II of Figure 8). This difference is shown by the higher price and by Q_L being less than MES. This added cost is not just a dead loss, for consumers benefit from the extra price they pay. For example, the local grocery store may charge higher prices, but to its neighborhood customers, the extra convenience can be worth the extra cost of shopping there. Or perhaps brand preferences are at work. For example, suppose that a restaurant has a better menu than the other restaurants in town. Some customers will be willing to pay more for these meals, along the demand curve shown in Figure 8. They pay a higher price, but they also get meals that they like better than those in the other restaurants.

The *second* deviation is idle capacity. Because output Q_L is less than MES, some of the firm's capacity (the amount $MES - Q_L$) stands idle most of the time. In practical terms, most retail shops have near-empty aisles for much of every day; most restaurants would like more customers than they have.

You can see these two distinctive features—idle capacity and extra pricing "for convenience"—in many stores that you deal with. Monopolistic competition is a special analytical case, but a familiar phenomenon in many day-to-day businesses.

Patterns and trends in real markets

You can now interpret the main patterns and trends of industrial structure in the U.S. economy. Recall the examples in Table 2, which were merely a selection to illustrate the main types of markets. Now it is time to consider the whole patterns and trends of industrial structure in the U.S. economy.

There are three main kinds of conditions to consider: (1) the aggregate share of total output produced by the largest firms in the economy (such as the share of the 50, 100 or 200 largest firms); (2) the degree of concentration in individual markets; and (3) conglomerate firms that operate in many markets.

Aggregate concentration

Aggregate concentration is the share of national output produced by the largest firms in the economy. All of the biggest firms are added together, to see how much of all U.S. economic activity is held in a few hands. This kind of concentration has little to do with competition, for it merely

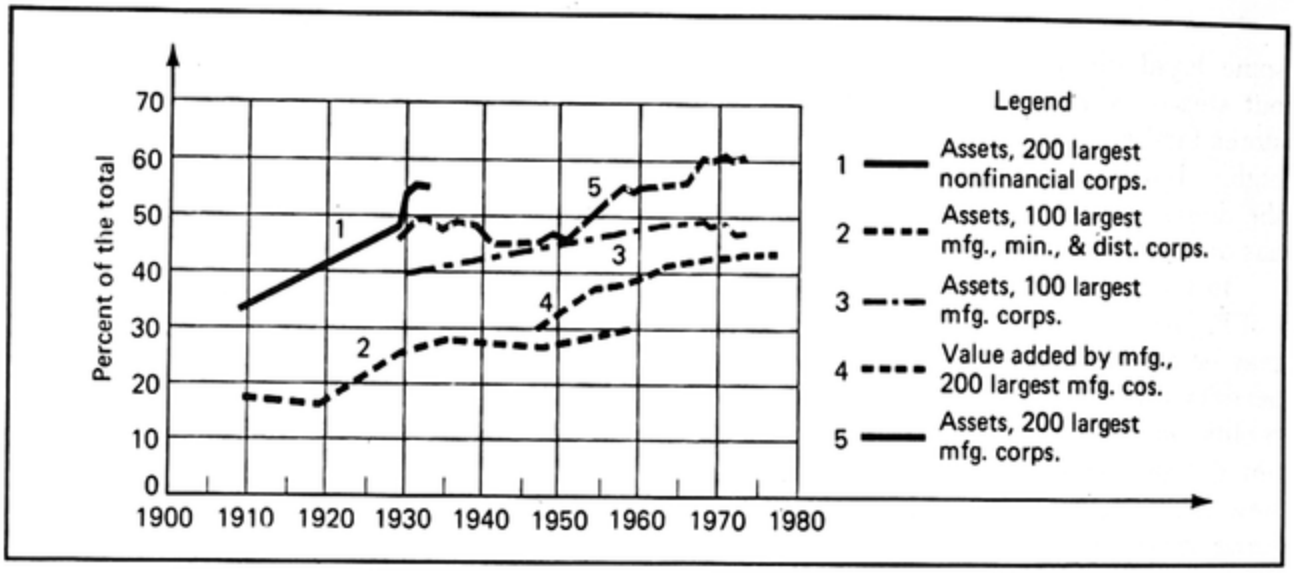


Figure 9 The scope and trends of large U.S. corporations

One can measure the largest firms' share of total assets or value added in the manufacturing and financial sectors. The trends have been up, though perhaps tapering off in recent years, and possibly even declining.

Source: Shepherd, *The Economics of Industrial Organization*, 1979, p. 114, and U.S. Census Bureau data.

totals up all kinds of companies, most of which do not compete with one another. If their total share is high, then big business may exert power over the economy. If their share has been rising, that might show that corporate power and control have been increasing in the entire economy.

The actual levels and trends of the largest 100 or 200 *industrial* firms are shown in Figure 9. (Data for the entire economy have not been prepared by economists, because there is no single correct measure to use.*) Large firms evidently hold substantial shares of total industrial assets and economic activity, but there is no spectacular dominance by just a few firms. The broad trend in the shares, which was upward after 1945, has been tapering off since the 1960s, perhaps even declining in the 1970s.

*Thus it is not technically meaningful to add up the assets of manufacturing companies with the paper assets of banks and insurance companies. If one added up all the companies' sales together, then retailing firms would be overemphasized because they have large sales volume but little capital or employment.

Concentration in individual markets

Because there are thousands of individual markets, which range from wheat and coal to steel, automobiles, newspapers, telephone service, banking, and restaurants, and the evidence about them is incomplete, economists cannot be precise about the degree of monopoly in the whole economy. Yet, close study over several decades has clarified the main patterns.

The Census Bureau defines about 440 manufacturing industries in the United States, and it reports their four-firm concentration ratios about every fourth year. Using these and other sources, a recent study has classified U.S. markets into four categories: pure monopoly, dominant firm, tight oligopoly, and all others.* The "all others" category includes loose oligopoly, monopolistic competition, and pure competition, all of which together can usually be regarded as very close to the fully competitive situation.

*William G. Shepherd, "Causes of Increased Competition in the U.S. Economy, 1939–1980," *Review of Economics and Statistics* (to be published 1983).

To test the trends over time, the study made estimates for three widely spaced years, 1938, 1958, and 1980. These estimates are summarized in Table 5 and Figure 10.

There are two main lessons from this study: ***First,*** the economy contains a wide variety of market conditions. ***Second,*** there has been a marked rise in competition in the economy. Pure monopolies were about 6 percent of national economic activity in 1939, but their share had shrunk to 2 percent by 1980. Dominant firms shrank almost as sharply, from 5 percent to a 2 percent share of national income in 1980. Tight oligopolies had no less than 36 percent of national income in 1939 and 1958, but then shrank sharply to 19 percent in 1980.

Taken altogether, these three categories of market power have decreased dra-

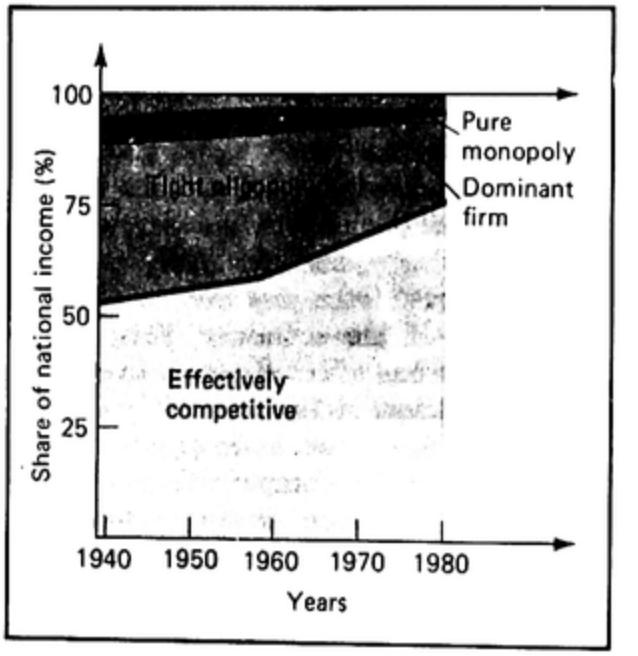


Figure 10 The trend of competition in the U.S. economy 1939 to 1950
In 1939, just over half of all national income arose in markets that were effectively competitive (they were either loose oligopoly, monopolistic competition, or pure competition). The competitive share rose to 58 percent in 1958, and then to 76 percent in 1980. Much of the rise occurred in the 1970s.

Table 5 Trends of competition in the U.S. economy, 1939–1980

Sectors of the Economy	National Income in Each Sector 1980 ($ billion)	The Share of Each Sector That Was Effectively Competitive 1939 (%)	1958 (%)	1980 (%)
Agriculture, forestry and fisheries	55	91.6	85.0	86.4
Mining	25	87.1	92.2	95.8
Construction	88	27.9	55.9	80.2
Manufacturing	460	51.5	55.9	67.9
Transportation and public utilities	164	8.7	26.1	39.1
Wholesale and retail trade	262	57.8	60.5	93.4
Finance, insurance, and real estate	211	61.5	63.8	94.1
Services	245	53.9	54.3	77.9
Totals	$1,510	52.4	56.4	76.3
The Share of Each Category in Total National Income				
1. Pure monopoly	38	6.2	3.1	2.5
2. Dominant firm	36	5.0	5.0	2.8
3. Tight oligopoly	283	36.4	35.6	18.0
4. Others: effectively competitive	1,153	52.4	56.3	76.7
Total	$1,510	100.0	100.0	100.0

Note: Totals may not match because of rounding.

matically. Production by monopolies, dominant firms, and tight oligopolies shrank from nearly half of national income in 1939 to less than one quarter in 1980. Correspondingly, the effectively competitive share rose from just over one half to 76 percent of the economy. This rise in competition has affected nearly every sector, as is evident in Table 5. As you can see from Table 6, the two main causes for this sharp rise in the competitiveness of the economy have been *import competition* and *policy actions* (to be explained in Chapter 12). These causes have acted with special force in the 1970s. Some major industries, however, have retained a high degree of monopoly.

Altogether, the shift toward increased competition has affected most of the economy and appears unlikely to be reversed. Much of it came unexpectedly in the 1970s, after decades in which U.S. industrial structure had gradually grown more monopolistic and more stable.

Conglomerate firms

Conglomerate firms operate in *many* markets, rather than in just one. They range from firms with two product lines to highly diversified enterprises with hundreds of branches and thousands of products. Such diversified enterprises have long existed. The British East India Company in the 17th to 19th centuries, for example, was a large and powerful conglomerate, operating in cloth, spices, minerals, and many other products. Modern conglomerates are sometimes quite small. But some of them are very large and own a string of subsidiaries with high market shares. ITT is one such conglomerate; among other activities, it owns Continental Baking ("Wonder Bread"), lumber, Avis (car rental), and Sheraton Hotels. RCA—with TV-set production, Hertz car rental, NBC, and many other lines—is a second example.

In fact, virtually all large firms are diversified to some degree, making a main

Table 6 Leading cases of rising competition in the U.S. economy

1.	*From Imports and Foreign Competitors*	
	Automobiles	Oceanic fishing
	Steel and products	Motorcycles
	Cameras	Television sets
	Copiers	Tires
	Shipbuilding	Aircraft
	Typewriters	
2.	*From Deregulation*	
	Airlines	Radio broadcasting
	Air freight	Long-distance telephone service
	Railroads	Banking
	Stock markets	Bus travel
	Some telephone equipment	Trucking
	Television broadcasting	
3.	*From Antitrust Actions*	
	Aluminum	Automobile rentals
	Metal cans	Certain professions
	Film processing	Some professional sports
	Motion pictures and theaters	Electrical equipment
	Cement	Telephone equipment
		Photographic supplies

line of products but also stretching into other fields. Though it is not usually regarded as a conglomerate, General Motors makes automobiles *and* buses, locomotives, trucks, and scores of other products. Many chemical and food companies are highly diversified within their broad sectors. Thus, the DuPont company makes thousands of chemicals; General Foods sells hundreds of disparate food products; and General Electric makes many types of electrical equipment.

Some conglomerates are only holding companies, which act mainly as financial supervisors and sources of capital for the operating firms they own. At the other extreme, many conglomerates exert tight management control over all of the detailed production activities of all their subsidiaries.

The distinction between *established* diversified firms and *new* conglomerates is also important. The older diversified firms were created decades ago and have become part of the established industrial structure: blue-chip firms that are accepted by bankers and other large firms. New conglomerates, in contrast, are regarded as upstarts and threats, for they often try to take over solid blue-chip companies against their will. Since the business press usually reflects the interests of the old-line firms, it often portrays the new conglomerates as dangerous and unreliable hucksters.

Yet, these newcomers often provide the new blood that is needed to stir up the staid ways of the business establishment. Scores of conglomerates were formed by hundreds of rapid mergers during the "go-go" stock market boom of the 1960s. Many of these soon fell apart, performed poorly, and were dismantled by the late 1970s. Yet others have been highly efficient and innovative.

Conglomerates are mainly neutral in their impact on competition and performance. One great economic benefit is that conglomerates often "take over" sluggish firms and then improve their efficiency. Just the threat of being taken over against their wishes will keep many firms' managers working harder to be efficient, so as to avoid an actual takeover. Since the managers naturally fear losing their jobs after a takeover, they often put up stiff resistance to the merger.

Each month you can find several dramatic takeover attempts reported in the financial press. The target company may escape, but even so, the trauma will usually change its attitudes drastically. Henceforth, the firm will tend to operate more efficiently, so that no other firm is tempted to take it over and make it even more profitable.

On the other hand, three harmful economic results can be caused by conglomerates. *First,* conglomerates often make mistakes, bungling their mergers and supervising their subsidiaries poorly. For example, the Penn-Central Railroad merger in 1970 was a spectacular failure, partly because the officers were busy trying to juggle real estate and other diversified projects. Whole trains got lost from time to time, and operations became confused. This type of wreckage affected scores of firms in the 1970s, and new fiascos still occur frequently.

Second, many conglomerate firms are distant absentee owners of their far-flung operations, with little local awareness or sense of responsibility. A fine local business, once it is bought out by a conglomerate based in a distant metropolis, is often shut down or moved away with little concern for how these decisions will affect the local community.

Third, some large conglomerates can exert power on local governments to get tax and other privileges. The firms can negotiate to pay lower taxes or obtain subsidies, under a threat to move their opera-

tions elsewhere. When the firms operate in many international markets, they can often exert the same pressures on small countries.

Yet, on the whole, most conglomerates are neither so sinister nor so virtuous as is often claimed in popular discussion.

Summary

This chapter covers the middle range of industry structure lying between pure monopoly and pure competition. The main points in the chapter are summarized below.

1. Degrees of competition can be divided into three major categories: dominant firms, oligopoly, and monopolistic competition.
2. A firm is said to be a dominant firm when it has over one half of the sales in its market and is more than twice the size of the next largest firm. Its demand curve is close to having the position of the entire market demand curve.
3. A dominant firm's position is usually protected by *entry barriers* that make it difficult for new rivals to enter the market. Such barriers might be due to heavily advertised *product differentiation,* which weds consumers to familiar products, or to substantial *economies of scale,* which require large amounts of capital and a large market share for efficient production. Other causes of dominance are *mergers,* and *patents, and trademarks.*
4. *Oligopoly* refers to an industry dominated by two or more rivals of approximately equal size. The key to oligopoly is *interdependence.* Each firm is aware that the actions of other firms will have an impact on its own market position. Oligopoly is usually measured by the *concentration ratio,* which represents the percentage of sales accounted for by the largest firms in the industry, usually the top four firms.
5. Oligopoly embraces such a wide variety of structures that no one "model" of oligopoly has been developed. Some of the general lessons regarding oligopolists' behavior are:
   a. Oligopoly has conflicting incentives, to both compete and collude. Collusion is more likely when firms have similar demand and/or cost conditions, when firms' managers are familiar with rivals' behavior, and when there is a high degree of concentration.
   b. Types of collusion range from tight and explicit forms, such as price fixing, to more informal, looser arrangements, such as price leadership.
   c. In oligopoly markets, there is a tendency for firms to converge not only on identical prices but on similar product features as well.
   d. In oligopoly, each firm's choices depend on what it expects the others to do. The oligopoly must therefore develop a *strategy,* to anticipate its rivals' reactions to its own actions.
   e. Because of the uncertainty arising from the firms' interdependence, oligopoly tends toward rigid prices, changed relatively infrequently. One theory developed to explain these rigid prices is the theory of the *kinked demand curve.* The kink arises from the fact that the firm expects to have an elastic demand if it increases its price (because other firms will not increase

their prices) and a relatively inelastic demand schedule if it lowers its price (because other firms will lower theirs).

f. Studies show that economies of scale are not a universal cause for oligopoly. There are many cases where oligopolistic firms have a larger market share than required for efficiency.

6. *Monopolistic competition* refers to markets in which firms have a small market share (less than 10 percent) and a differentiated product that gives the firms' demand schedules a slight downward slope. Freedom of entry and exit exists for these markets and there is no recognized interdependence among the firms.
7. Because there is freedom of entry and exit, a monopolistic competitor may earn only a normal return (economic profits = 0) in the long run, as was true for perfect competition. Under monopolistic competition, however, average total cost and equilibrium price will be higher and quantity lower than under perfect competition.
8. Studies of the U.S. economy show that there has been a noticeable rise in the degree of competition in the economy during the 1938–1980 period. The cause of this has been both an increase in import competition and policy actions, including antitrust actions and deregulation.
9. Conglomerate firms, which operate in more than one market, have become an increasingly familiar part of the economy since the 1960s. Although they can produce harmful economic results, such firms seem, in general, to be fairly neutral in their impact on industry competition and performance.

Key concepts

Dominant firms
Oligopoly
Concentration
Tight oligopoly
Loose oligopoly
Interdependence
Price fixing
Monopolistic competition
Conglomerate firms

Questions for review

1. Given what you know of their markets, classify the following firms as dominant firms, oligopolies, or monopolistic competitors. State the reasons for your decisions.
   a. U.S. Steel
   b. Local family-owned restaurant
   c. Procter & Gamble
   d. Ford
   e. Kodak (film market)
   f. The largest local newspaper
   g. Budweiser
   h. Campus-area movie theater
2. Consider the following statement and determine whether it is true or false. Explain your answer.

   Even if the technology of a certain industry involves substantial economies of scale, new firms need not be at a competitive disadvantage. After all, both old *and* new firms would need to reach a high level of output to achieve minimum efficient scale.
3. Classify each of the following statements as true or false. Explain your answers.
   a. Firms in an oligopoly will have no incentive to break a price-fixing agreement as long as the agree-

ment results in higher profits than each firm would have in the absence of any joint action.

b. According to the theory of the kinked demand curve, an oligopoly with a pessimistic view would expect to face an inelastic demand if it increased its price and an elastic demand if it lowered its price.

4. As its name implies, monopolistic competition is an industry structure that shares characteristics of both monopoly and competition. In what ways is monopolistic competition similar to competition? In what ways is it similar to monopoly?

5. Explain how the combination of the *characteristics* of monopolistic competition and the general rules for profit maximization will lead to the existence of idle capacity in monopolistic competition.

6. Consumers pay higher prices under monopolistic competition than they would if the industry were competitive. Yet, society gets something in exchange for these higher costs and prices. Explain.

7. Many judge that the existence of conglomerate firms is close to neutral in terms of impact on industry performance. What are two of the possible benefits and two of the possible harmful effects of conglomerates?

12

Policies Toward Monopoly Power: Antitrust

As you read and study this chapter, you will learn:

- the origins and standards of antitrust policy
- the economic elements of antitrust cases
- the three specific parts of antitrust: toward existing concentration, mergers, and price fixing and other actions

When European economists discuss what is special about the United States economy, they frequently mention its large size and abundant natural resources. But even more distinctive than these, they often say, is "your touching faith in competition, as shown by your antitrust policies." They also point to our treatment of the main public utility industries, where we regulate private companies rather than convert them to public ownership. Indeed, antitrust and regulation are unique American experiments, and they are the country's main defenses against monopoly power.

Antitrust policies promote competition throughout most sectors of the economy. Price fixers are punished and anticompetitive mergers are stopped. Regulation sets limits on utility prices. These policies are the focus of intensive debate because they deal with urgent, complex problems whose stakes run into many billions of dollars.

To their critics, U.S. antitrust and regulation policies appear to be weak, deceptive, or even harmful. However one judges these policies, they deserve to be studied closely, for they can decisively affect the nature of the competitive, market-based economy.

After briefly summarizing the origins and economic standards of antitrust policy in the first section, we devote the rest of this chapter to discussing antitrust policy. The main lines of actual antitrust policies are discussed in the second section. The third section considers the three specific parts of antitrust: toward existing concentration, mergers, and price fixing. Throughout, we have space to present only the basic patterns of this complicated subject.

Origins and standards of U.S. antitrust policies

Before presenting U.S. antitrust policies one by one, we need to review their origins so that you will better understand their nature. The policies have been hammered out over decades of turbulent political action, and they remain the focus of intense battles. With such mixed origins and continuing pressures, the three policies—antitrust, regulation, and public enterprise—are naturally imperfect rather than ideal. The task is to discover what these policies are really doing to the economy.

Three waves

There have been three major waves of policies toward business. The first came in 1885–1915, as antitrust policies and regulatory agencies were begun. Then, during 1933–1950, a second wave occurred, especially as airlines, telephones, and electricity came under regulation. Finally, in 1965–1975, the third wave created a battery of agencies regulating safety and health. Figure 1 includes some of the main antitrust and regulatory parts of these waves.

Each wave reflected the public's discontent with recent business performance and its belief that new actions were needed. Many of the actions were inadequate, went too far, or applied the wrong incentives. Some of the faults were corrected after each wave, when efforts were made to trim back the policies.

The most crucial formative period for policies was 1890–1910. Before 1890, there had been a scattering of rules and laws dealing with the early forms of business, mostly at the local and state level. In cities, the gas and water utilities were controlled in various ways, often under city ownership. The charges for using turnpikes and canals presented problems of monopoly pricing, met in diverse ways by the various states.

Upon this localized scene the U.S. industrial revolution burst with great force during 1865–1900. The railroads spread across the country, forming monopolies in some regions and charging discriminatory prices ("what the traffic would bear"). Stirred by the Civil War and the railroad boom, heavy industries grew rapidly. Gold rushes, land rushes, the invention of electric light systems and telephones, the dramatic growth of the oil industry during 1870–1890—these and other new developments created an industrial transformation of the growing country. Moreover, the high financiers—especially J. P. Morgan—were busy forming "trusts" in many industries by merging scores of little firms into big ones.

The 1880s brought rising public agitation about these changes. Farmers organized to fight price gouging by the railroad monopolies. They and other citizens increasingly denounced the new industrial trusts. Amid sharp political debates, there

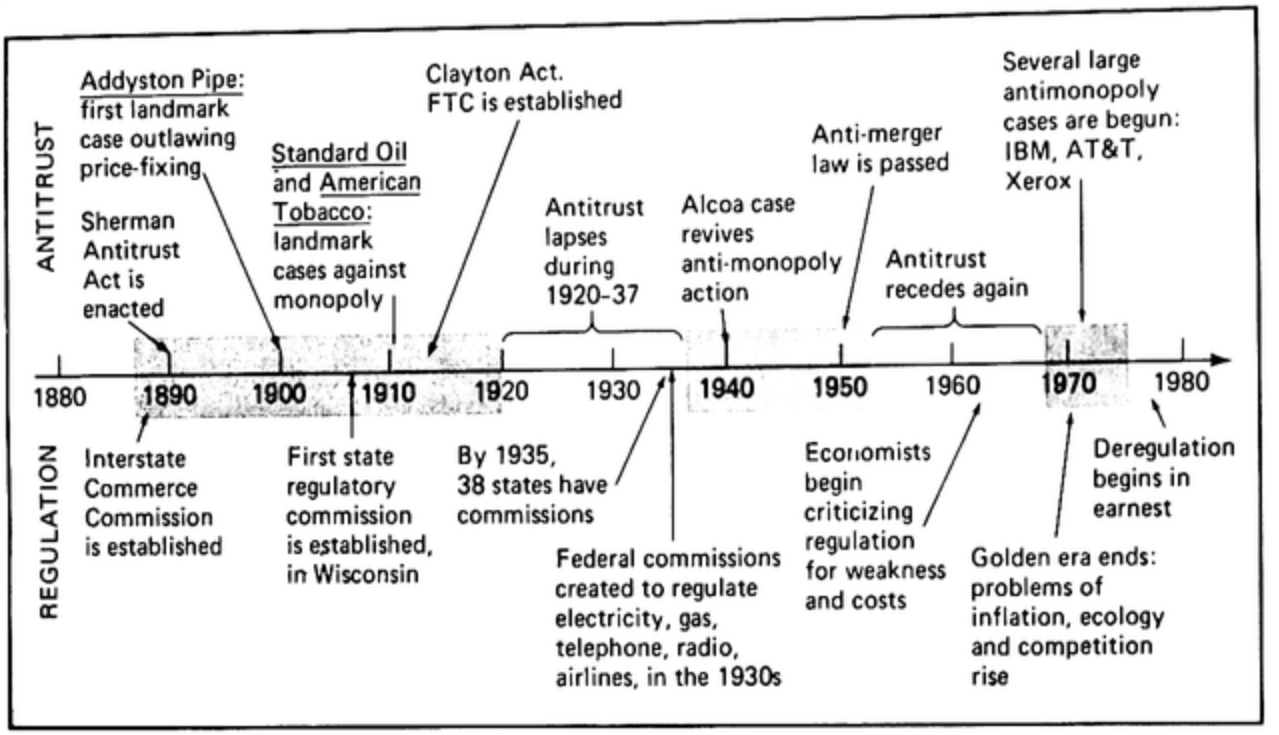


Figure 1 Main events in antitrust and regulation

followed two kinds of policy action. *First*, regulation was established, starting with the Interstate Commerce Commission in 1887 to regulate the railroads. Then state regulatory commissions were created from 1907 on, to regulate electricity, telephones, and railroad traffic within the states. These steps accepted the *private* ownership of the basic utility operations, thereby turning against the public ownership of the railroads and city utilities that was already common in other countries.

The regulatory commissions were supposed to control the private utilities strictly, permitting only "fair" rates of return and "just and reasonable" prices (i.e., without too much price discrimination). By the 1930s, most of the states had their own regulatory commissions, and new federal commissions had been created to cover *interstate* operations in the main utility sectors.

Second, ***antitrust policy*** was created in 1890 to reduce monopoly in the rest of the economy. It was called "antitrust" because it was aimed against the creation of industrial trusts (which were combinations of firms to fix prices). The Sherman Act of 1890 outlawed both monopolizing by one firm and collusion among competitors. After some delay, the law was applied firmly to price fixing in 1899 and to monopolies in 1911. As further enlarged in 1914 and 1949, the U.S. antitrust laws became a uniquely thorough method for stopping industrial monopoly and price fixing.

Yet, antitrust and regulation have had checkered careers since 1920, reflecting the larger economic trends and political swings. Antitrust has veered between great waves of action (in 1938–1952 and 1968–1980) and relatively inactive periods (the 1920s and 1952–1968). Regulation took a long time to get established, with full coverage and powers being reached only in the 1950s. At most times, both antitrust and regulation have been sharply criticized, by business for being too harsh and "antibusiness," and by others for being too weak and "probusiness."

The selections of people to run the agencies are made under political pressure, often resulting in mediocre appointments. The agencies' budgets are also political decisions. Often the strongest

lobbying is done by the very companies that the policies are supposed to control. In these and other ways, the companies may influence policy as much as policy influences the companies' actions.

That is what is meant by those two-way arrows in the middle of Figure 1, Chapter 11. Actual policies evolve in a rugged political setting, often under intense pressures from many sides. Therefore, one must expect antitrust, regulation, and public enterprise to be imperfect and limited. We will show, nonetheless, where they have often come close to sound economic results.

Public enterprise has not been fully eclipsed. It has continued in the U.S. Postal Service, many hundreds of city electric and transit systems, thousands of city water works, and still other cases to be noted in Chapter 13. Yet, it has scarcely been tried at all in manufacturing and finance. This is a sharp difference between the United States and many other countries. Since 1945, foreign experience with public enterprise has included all economic sectors.

Standards of efficient policies

Economists judge policies by a simple standard called "cost-benefit analysis." It is merely a specialized version of the fundamental economic comparison we have pointed out repeatedly in earlier chapters. Recall that each economic action provides benefits, usually in the form of a good that people will pay for. The same action will also incur costs—the effort and resources needed to produce the good. Efficient economic decisions will carry production until the marginal benefits just equal marginal costs, as we stressed in Chapter 9. The marginal unit is just worth its cost.

Public policy choices face exactly the same criterion of efficiency: ***Each action should be carried out up to the level where its marginal benefits just equal its costs.*** But here the benefits are the *public's* benefits that arise when monopoly's effects are prevented: greater efficiency, lower prices, more rapid innovation, and other desirable goals. If the action is wise, then the public reaps these benefits through the improved performance of the economy.

But there are also policy costs to consider. They are mainly incurred by the public in paying for the agency to take the action. Agencies use resources (in staff members' salaries and other costs) in applying their policies. For example, the antitrust chief considers whether to prosecute five bakers in St. Louis for fixing the prices of their bread. The case will cost $10,000 to carry out (in working time, travel, etc.), and it will only improve conditions in one small market. Ideally, the official weighs the benefits and costs, at least approximately, in such a marginal case. In practice, the judgments may be faulty or the agencies may be given either too many or too few resources.

This kind of cost-benefit comparison is the correct basis for appraising antitrust, regulation, and public enterprise. We will show what the agencies have been doing, which is fascinating in itself and full of human drama. But the ultimate economic task is to judge which policies have been carried too far or not far enough, in light of their probable costs and benefits. The specific sums are rarely easy to measure, so that careful judgments have to rely on reasonable likelihoods rather than actual numbers. For example, was the recent large antitrust case against AT&T worth its cost? Should the regulation of long-distance telephone service, railroads, banking, even of electric service, be withdrawn? In these and scores of other cases, the student can practice the kind of cost-benefit thinking that is the test for rational policies.

U.S. antitrust: Forms and coverage

Antitrust policies deal with anticompetitive conditions that arise in the broad range of ordinary markets. As we showed in the preceding chapter, firms are always tempted to eliminate the annoyances of competition by merging with each other, fixing prices, or other tactics. Competition often needs a helping hand, or even sharp antitrust actions, to be effective.

Since 1890, it has been illegal to monopolize markets, or to fix prices, or to deal abusively with one's competitors. These antitrust laws seek to keep competition effective, so that markets will be both efficient and fair. Like any laws, antitrust laws have to be enforced, by prosecuting and penalizing those who violate them. The laws and the enforcement are quite imperfect in practice, as we will shortly show. Yet for all its faults, American antitrust policy stands out as the world's great policy experiment in trying to maintain competitive markets.

U.S. antitrust policies offer fascinating lessons about the economics of competition. Our aim in this chapter is not to convey the details of policy; they are often minor and confusing, and they change frequently. Rather, we wish you to learn how antitrust policies produce economic results. Then you can judge the value of competition and of antitrust itself.

Antitrust policy is pervaded by controversy, and many of the issues are debated intensely. The policy toward price fixing has always been strict; the treatment of mergers has been mixed; while the treatment of dominant firms and tight oligopoly has mostly been mild. Are these policies right, or should they be changed? If changed, should it be toward stricter or gentler enforcement? You will need to keep an open mind and to think independently about these issues. Basic economic concepts can clarify the main points of the debates, but they do not lead to definitive answers.

The agencies and laws

Antitrust policy consists of (1) *agencies* that enforce (2) *laws*. There are two enforcement agencies: (1) the Antitrust Division of the Department of Justice; and (2) the Federal Trade Commission, an independent agency created in 1914. Until the late 1930s, they were tiny agencies with minuscule budgets and a few score staff members. During the second policy wave, they grew to about 300 lawyers each by 1950, and then stabilized. They expanded again after 1970, during the third policy wave. Their budgets for antitrust enforcement were $4 million each in 1950, and still below $12 million in 1970. By 1981, their budgets for antitrust enforcement had grown to about $40 million each. Yet they were still tiny compared to an economy of $2 trillion and a total federal budget over $600 billion.

These resources are thinly spread. Many major industries are dealt with by just a few lawyers and economists. Many other sectors, especially new industries, are given only passing attention. A single big case can engross a sizable share of the whole agency's resources. At the top, the agencies are run by political appointees, who are usually in office for only three years or less. Most significant actions take between five to fifteen years to run their course. Therefore, policies often lack sustained guidance.

The setting for the agencies includes (1) the rest of the government and (2) private antitrust resources. Though they are nominally free from outside interference, the agencies are subject to various pressures. Their budgets are settled by the executive branch and Congress. Firms try to

use officials in the executive branch (in the White House, Defense Department, and elsewhere) and members of Congress to influence the agencies. Actions can be appealed to the appellate courts and the U.S. Supreme Court, either to reverse decisions or merely to delay the process.

On the private side, the defendant firms often deploy large resources to resist or manipulate the policy efforts. The private antitrust bar includes about 10,000 lawyers. Large firms can routinely apply 5, 10, or 20 times as many lawyers and experts to a case as the public agencies can. This fits their large stakes in the outcomes, often running into hundreds of millions of dollars. They may spend up to the total amount of profit that is at stake in order to win the case.

Private antitrust suits—by one firm against another—often trigger or supplement actions by the public agencies. Each year there are over a thousand such private cases, in a great variety of markets. In theory, they should neatly fill in any gaps in public policies. In practice, private cases often are lacking precisely where they are most needed.

Because they are so small, the two antitrust agencies mainly try to develop a series of precedent-setting cases, rather than to pursue and catch every firm that might be breaking the antitrust laws. We will present some of those landmark cases in the third main section of this chapter.

The laws are broad and powerful, so that even the two tiny agencies can have some strong effects. The Sherman Antitrust Act, passed in 1890, is the country's basic antitrust law. It has two main sections, which are summarized in Table 1. Section 1 outlaws conspiracies to "restrain trade." It mainly applies to explicit price fixing, which was discussed in the preceding chapter. Section 2 outlaws "monopolizing" or "attempts to monopolize"; that is, firms that clearly try to dominate their markets.

Table 1 The basic U.S. antitrust laws

1. **Restraint of Trade** (Sherman Act, Section 1)
   Collusive actions, such as price-fixing, market-rigging, and sales-allocating schemes, and other restrictive actions, are all forbidden.
2. **Monopolizing** (Sherman Act, Section 2)
   Both monopolizing and *attempting to monopolize* a market are illegal.
3. **Mergers** (Clayton Act, Section 7, amended in 1949)
   Any merger that may substantially reduce competition in any market is illegal.
4. **Other Actions That Are Prohibited:**
   a. Interlocking directorates (one person serving on the boards of two competing companies).
   b. Price discrimination that harms competition (Robinson-Patman Act of 1936).
   c. Exclusive and tying contracts (Clayton, Section 3). (If good A can only be bought by also buying good B, then the two goods are "tied.")
   d. "Unfair" methods of competition (FTC Act, Section 5). These are unspecified in the law but would include abusive or extreme actions.

History

After a shaky start, the Sherman Act was applied firmly against price fixing in 1897 in the *Addyston Pipe* case (the case is summarized in the third main section of this chapter). That strict prohibition of price fixing is still enforced. No longer able to collude with each other after 1897 because of this new precedent, many firms simply merged. Figure 2 charts the waves of mergers and antitrust actions since 1890. The pattern of action and response is clear. Each new wave of mergers stirred anxiety in the populace that corporate power was being enlarged. Antitrust officials then renewed their efforts.

First came Theodore Roosevelt's "trust-busting" campaign, actually carried

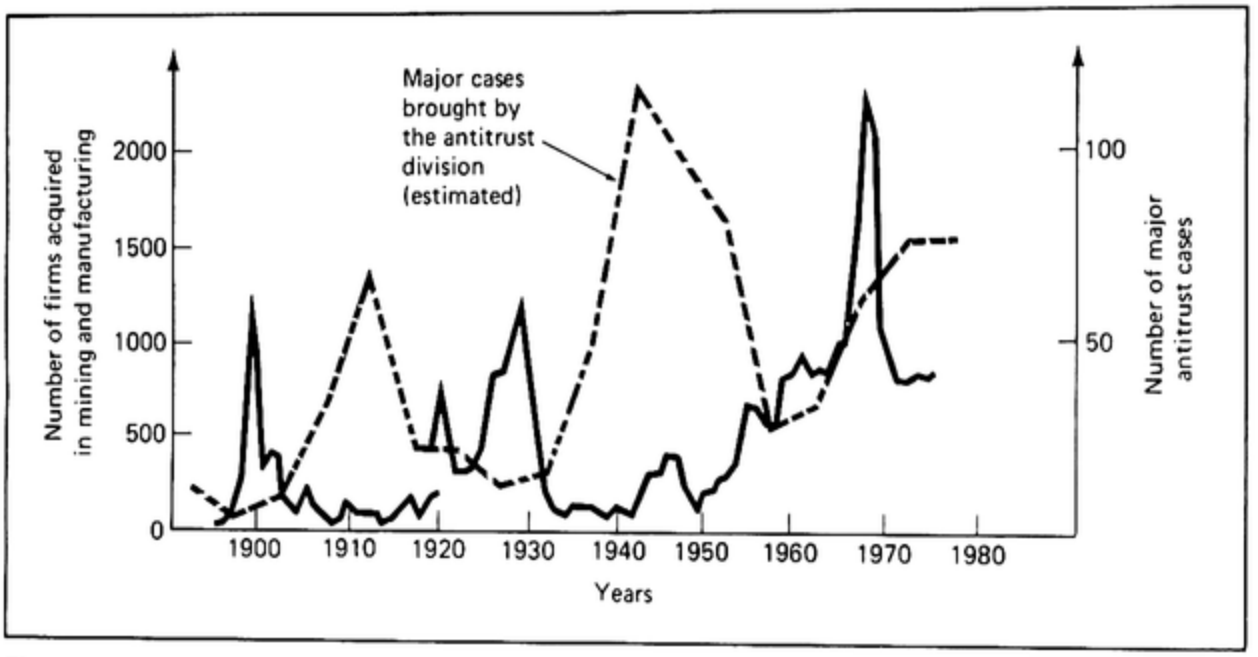


Figure 2 The pulses of merger and antitrust activity

The three merger waves—of 1897–1901, the 1920s, and the 1960s—have been dramatic and turbulent. Antitrust actions have also come in three distinct waves, as shown by the number of cases (weighted for importance). To some extent, the actions have been a response to the mergers and other industrial events.

Source: Mergers: W.G. Shepherd, *The Economics of Industrial Organization* (Englewood Cliffs, N.J.: Prentice-Hall, 1979). Antitrust adapted from Richard A. Posner, "A Statistical Study of Antitrust Enforcement," *Journal of Law and Economics*, 13 (October, 1970), pp. 374–381.

through mainly during William H. Taft's presidency in 1909–1913. The Standard Oil Company and the American Tobacco Company were each divided into several companies, and several other firms were required to sell off a plant. For example, Standard Oil was separated back into its original regional monopolies, while American Tobacco was divided into three firms in 1913, which are still operating—American Tobacco, Liggett and Myers, and P. Lorillard.

Because the Sherman Act is so terse and broad, business interests soon demanded more details about exactly which actions were illegal. Under Woodrow Wilson's "progressive" approach, the Clayton Act in 1914 was written to try to cover those details, though only a few specific offenses were spelled out. The Federal Trade Commission was also created at that time to enforce the law.

Antitrust then went into hibernation during the "Roaring Twenties," even as the second great merger boom mounted to its peak in 1929. The balance of antitrust actions actually favored mergers and cooperation among firms. But the Crash of 1929 and the Great Depression of the 1930s renewed pressure to act against what was seen as rising corporate power. The second deep antitrust wave, from 1938 to 1952, touched many dominant firms. Thus, the aluminum monopoly (Alcoa: the *Al*uminum *Co*mpany of *A*merica) was challenged in 1937, and two new competitors were finally added in 1950. The monopoly in metal cans (American Can Company and National Can Company) was subjected to restraints by a court decree in 1950. The dominant National Broadcasting Company (NBC) was required to sell one of its two radio networks in 1943, thereby creating the American Broadcasting Company

(ABC). Movie companies were forced by the *Paramount Pictures* decision in 1948 to sell their movie-theater chains. There were also major cases against some tight oligopolies, including the leading cigarette companies.

The pace of antitrust actions slowed again from 1953 to 1968. The only exception was the application in 1962 of strict new limits against mergers among competitors. On other antitrust fronts, little was done to reduce market power or to stem conglomerate mergers during the "go-go" merger boom of the 1960s.

After 1968, the antitrust pulse quickened, especially with several large cases alleging that IBM, Xerox, AT&T, and the big cereal companies had monopolized their markets. Then in 1981, the Reagan administration reduced the reach of both antitrust agencies. Efforts to lessen market dominance were trimmed, especially at the FTC. The AT&T case was settled and the IBM case was dropped. Merger rules were relaxed, resulting immediately in a merger boom in 1981. Only price fixing was to be pursued as strictly as before.

After all these fluctuations, antitrust enforcement has settled into a mature state of moderation, part of the industrial and social fabric, but beset by sharp debates about its value and wisdom. Conservatives regard antitrust as mainly a harassment of businesses, doing little good in an economy that is already strongly competitive. By contrast, many liberal experts describe antitrust as weak and poorly managed, a token effort that leaves much monopoly power untouched.

These are old, old issues, always lively and sharply argued. They are also big-league issues because billions of dollars and high degrees of innovation are often at stake. The 10,000 antitrust lawyers are continually engaged in studying and litigating, and at any time there are many hundreds of cases in progress. Many of those cases are brought by private firms, alleging that they have suffered from antitrust violations by other private firms (by being overcharged for goods, by being driven out of business, etc.). Often a firm's fate is affected more deeply by an antitrust case than by any conditions out in the market itself. Almost every issue of the economics and legal journals—and of *Fortune, Forbes, Business Week,* and the *Wall Street Journal*—has articles discussing these topics. The conflicts of evidence and views are often dramatic.

Antitrust criteria

An economist must try to see through the legal details of antitrust to discern its (1) economic criteria and (2) economic effects. Table 2 presents the main antitrust criteria that have evolved in the courts since 1890, in hundreds of precedent-making cases and decisions.

The wording of the statutes sounds broad and conclusive, but the laws have come to be applied only within "reasonable" limits—following what the lawyers call a "rule of reason." In practice, "reasonable" means what the courts will accept, and the limits of enforcement often move as the courts change and as the country's political momentum shifts. For example, the Supreme Court under Earl Warren in the 1960s tightened the criteria against mergers and dominant firms. The "Burger Court" of the 1970s and 1980s has drawn back the margin of antitrust enforcement, toward more moderate treatment of mergers and of the pricing tactics used by dominant firms.

Throughout, one pivotal issue has emerged: the definition of the market. *The two sides usually offer sharply differing definitions of the true extent of the market.* The plaintiff (an agency or a private company

Table 2 ***The main economic patterns of antitrust policy***

1. **Toward Existing Structure** (mainly Sherman Act, Section 2).

   Firms with market shares above 60 percent, high barriers to entry, and high rates of profit are to be sued. If market power is high and seems to have economic costs, the court then weighs possible economies of scale and the danger that an antitrust conviction will harm a broad range of the company's stockholders. Although its decision on the case will usually reflect all these factors, the court's formal opinion will usually cite only the market power and any abusive actions by the firm.

   If the firm is convicted, the remedies and penalties applied to it will usually be only moderate (such as levying fines or setting limits on the firm's future actions), rather than requiring the firm to sell part of its capacity.

2. **Toward Mergers** (Clayton Act, Section 7).

   *Horizontal mergers** of firms with a combined share above 10 percent of the market will usually be challenged by the Antitrust Division or the FTC. If the degree of concentration in the industry is rising, and if the merger would not give demonstrable economies of scale, the courts will usually stop the merger.

   *Vertical mergers** of firms with more than 10 percent each at their stage of the industry will usually be stopped.

   Conglomerate mergers will usually be permitted unless they unite two firms that are dominant in their own markets.

3. **Toward Price Fixing and Other Such Actions** (Sherman Act, Section 1).

   All efforts to fix prices will be prosecuted if they are discovered. This is true even if the firms have only a small share of the market. Most firms will be convicted, as long as there is tangible proof of the collusion. Most forms of market splitting (where firms divide up the sales in the market) and other explicit collusion will also be sued and convicted.

   Tacit collusion is exempt from prosecution, unless it becomes very thorough and obvious.

*Horizontal mergers are between competitors in a market (e.g., between General Motors and Ford). Vertical mergers are between firms at different stages in the chain of production (e.g., between a steel company and the firm supplying its iron ore).

claiming to be a victim of monopoly) urges a narrow definition, which gives the defendant firm a high market share. The defendant claims instead that the market is much larger, so that its share of that market is small. The court's decision on this point often governs the rest of the case, for if it accepts a large market, then harmful market power could not exist. Generally the Warren Court defined markets narrowly, whereas the Burger Court has accepted broad market definitions.

Precedents

Policy is applied by bringing individual cases, which the agency lawyers develop, bring to trial, and try to win. Some of the cases become vast, involving scores of lawyers, years of preparation, and millions of documents, as we will soon show. Many others are compact and clear, taking only a year or two from the violation to the final decision. The courts' decisions in these cases set precedents, which then govern subsequent cases. We will present various landmark cases in the third main section.

The precedents all reduce to fairly simple patterns, as shown in Table 2.

Toward existing structure The threshold criterion for prosecution has become (1) a 60 percent market share, (2) *plus* some evidence that the firm intended to gain dominance or has been unfair. Since any firm with a market share below 60 percent will almost certainly be acquitted, no cases are brought against firms in that range. Even with a market share above 60 percent, a firm can argue that its position arose from "superior skill, foresight, and industry" (that is the usual legal phrase), so that it deserves its dominance. Judges are often persuaded by that argument. IBM had particular success with it, winning a series of

cases since 1968. Market shares even above 60 percent often go untouched for decades because the antitrust officials expect that any suit against the firm would fail in the courts.

Mergers

HORIZONTAL MERGERS will usually be stopped if the resulting firm would have more than 5 percent of the market. Thus, a merger between Ford and Chrysler automobile companies (with market shares of 20 and 10 percent) would almost certainly be prevented. A merger of two 8 percent firms will be permitted if they can prove that they would thereby achieve economies of scale.

VERTICAL MERGERS will usually be stopped if the firms have more than 5 percent each of their markets. For example, Ford was stopped from buying Autolite, a firm that sells 20 percent of automobile batteries (mostly to the auto companies). Such a merger would have excluded other battery companies from a fair chance to sell batteries to Ford. However, Ford would be permitted to build up its own battery company, if it wished, by creating new capacity.

CONGLOMERATE MERGERS are left mostly untouched, especially if they involve small firms and small market shares. Thus, a tire company might buy a bakery, and then a railroad might buy the tire company. Since no market shares are increased in any market, an antitrust challenge to the merger would probably not occur. Only if each of the firms has a large share of its market *might* the merger be challenged.

Price fixing Price fixing is treated most strictly. The courts will not permit a defense of it on the claim that is was "reasonable." The agency merely needs to show at trial that the price fixing occurred, even without proof that its effects were strong. The courts will then usually convict it *per se* (that is, guilty "in itself").

PRICE DISCRIMINATION Little firms often accuse bigger ones of hurting them by price discrimination. The damage can occur if the larger firm makes selective price cuts, drawing customers away from an efficient small firm and thereby driving that firm out of business. Standard Oil, IBM, Xerox, and hundreds of other firms have been accused of this practice. But convictions on such charges are unsure, and the penalties are usually light.

Economic effects

The economic effects of these policies are debatable, but they probably have been as follows. A few major dominant firms have been reduced in size, mainly in actions taken before 1950. Tight oligopoly has not been touched, effectively challenged, or changed. Many horizontal mergers have been forestalled since 1962, though hundreds of mergers from 1890 to 1962 had already led to substantial concentration in many industries. Price fixing has mostly been driven underground, probably eliminating a large share of it. Yet, as we noted in the previous chapter, secret collusion does continue routinely in many industries and in a variety of tight oligopolies.

Altogether, antitrust policies have probably kept U.S. industrial concentration and the extent of price fixing much lower than they would otherwise have been. If antitrust were abolished, a large merger boom would immediately occur, raising concentration sharply in many industries. Formal price-fixing cartels, with official staffs and binding contracts preventing price competition, would be created in thousands of markets. Therefore, antitrust has created important economic

benefits, which continue quietly because antitrust itself continues.

Yet, the economic effects of antitrust have also caused some imbalance. Dominant firms are now largely free to set prices internally over large shares of the market. Thus, General Motors sets prices for some 45 percent of the U.S. automobile market and 85 percent of the locomotives market. But the little rivals of dominant firms (say, two firms with 6 percent market shares each) can neither meet to fix their prices nor merge with each other. Though they would affect only 12 percent of the market, they would be pounced upon by the antitrust agencies.

In cases like this, the law is gentle to the big and strict toward the small. Once a firm has gained dominance, it is largely immune from antitrust actions, free to do things internally that lesser firms cannot do among themselves. Ideally, antitrust would be equally strict toward dominant and little firms.

Antitrusts's ultimate effects are *debatable*, then, perhaps promoting efficiency on the whole, but perhaps also lacking balance. The effects are also *limited*, for antitrust reaches only to part of the economy. As Table 3 shows, most utilities, local markets, newspapers, professions such as the law and medicine, all labor unions, all patents, much weapons production, and most public enterprises are exempt from antitrust. The antitrust domain is, therefore, the core of national manufacturing industries and trade, altogether less than half of the U.S. economy. Meanwhile, various other policies directly reduce competition in many markets. They are summarized in Part 2 of Table 3. On the whole, antitrust's reach is far from complete. And even where it does reach, its resources are usually stretched thin.

Think of antitrust as *interacting with industry*, not standing above it exercising lordly powers. Like any other policing agency, antitrust officials are influenced by industry, by Congress, by the executive branch, and by swings of popular attitudes. Policy choices are often political, mistaken, rash, or too cautious: in short, thoroughly human, fallible, and changeable. Yet, the basic economic effects—against price fixing and mergers—remain relatively steady.

Table 3 Departures from antitrust: Exemptions and policies that reduce competition

1. **Exemptions**
   - Much local and statewide activity: construction, shops, repairs, services.
   - Labor unions at all levels.
   - Utilities and urban services: electricity, gas, telephone.
   - Social services and health services: schools, hospitals.
   - Public enterprises: many electric, transit, and water systems at the local and regional levels.
   - Farm and fishery cooperatives: dairy cooperatives.
   - Many military suppliers: aircraft, tanks, ships, ammunition, etc.
   - Baseball and, to a lesser degree, other professional sports.
   - Newspapers' joint publishing arrangements in many cities.
2. **Policies That Reduce Competition**
   - Tariffs and other barriers to international trade, such as quotas and agreements to limit imports.
   - Patents: they provide a monopoly for 17 years.
   - Banking regulation that prevents new entry in many banking markets.
   - Price raising for certain farm products (milk, tobacco, etc.) is enforced by the U.S. Department of Agriculture.
   - Shipbuilding and shipping: price fixing is enforced by the Federal Maritime Commission.

Specific parts of antitrust

Antitrust actions toward existing concentration
Defining the market The court decisions in cases involving an existing concentration of power usually hinge on the market

Is Antitrust Necessary?

Antitrust has severe critics, even apart from those businesses that would simply like to be free of its constraints. In recent decades, the attacks have come from two opposite sides.

Free-market liberals believe that nearly all markets are highly competitive. Only governments create lasting monopoly, they say. Any private market power will quickly be wiped out by new competition. Schumpeter's sequence of temporary monopolists (recall the box in Chapter 11) gives a similar result.

Accordingly, a minimal antitrust policy is needed, merely to stop blatant price fixing. Dominant firms, mergers, tacit collusion, price discrimination, and all other restrictions will not harm competition. Antitrust actions against them are useless or make matters worse.

Another group, whose most prominent spokesmen are John K. Galbraith and Lester Thurow, regards large firms as so efficient and innovative that efforts to divide or limit them are futile. If economies of scale and innovation are large, then antitrust should leave large firms untouched, even if their monopoly power is strong. Antitrust should also leave lesser firms free, for they are innocuous, and it is unfair to punish them for doing what large firms are free to do. If not abolished, antitrust should at least be pruned severely. Some degree of price controls or public ownership may be needed for the more powerful large firms.

These and other views all pivot on the same underlying economic conditions: the economies of scale, the effects of monopoly power, the sources of innovation, and the speed at which monopoly is eroded by market forces.

share of the leading firm. That, in turn, depends on the relationship between the firm's sales and the size of the market. Therefore, defining the market is a crucial step.

As Chapter 4 already noted, the market's true size depends on the extent of substitution among goods. That degree of substitutability is measured by the cross-elasticity of demand. We presented that concept in Chapter 4. Now we will show how the concept is applied step by step in antitrust decisions.

The relevant market The concept of "the" market rests on the choices made by consumers, as they compare substitutable goods. These choices are usually made along two main dimensions: the *kind* of good being exchanged (the "product type") and the *geographical area*. We will now consider those two dimensions in turn, just as the courts do in typical antitrust cases.

Consider *product types* first. The market should include those goods that are substituted freely. Thus, the newspaper market might include magazines and/or TV and radio broadcasts. But most buyers of newspapers will buy the newspaper even if its price rises (or if other goods' prices fall). Therefore, newspapers are in their own separate market because there does not appear to be much substitution between newspapers and television.

Likewise, the *geographical area* of the market will be defined by the choices that consumers make. Thus, most people just buy the local paper, while some also, or instead, buy the Chicago *Tribune* or New York *Times*. Yet very few Boston *Globe* readers will shift to the Cleveland *Plain Dealer* or Des Moines *Register*, even if the *Globe's* price triples. Therefore, the *Globe* is in a separate geographical market from the others, even if its product type (newspaper) is identical. In the last chapter, we presented a number of such local newspaper markets.

Both of these elements involve cross-elasticities of demand, but such elasticities are rarely calculated. A reasonable, careful judgment about the main scope of the market is usually all that the judges have to go on. In antitrust cases, the defendant firm invariably claims that the market is large (e.g., a defendant newspaper might try to include all newspapers in all cities within 200 miles, plus all magazines sold in that area). The plaintiff claims, to the contrary, that the real market is much narrower (e.g., solely the local newspaper market).

When the case has been argued and the evidence presented, the judges then have to decide what the "true" market is, as best they can. Any economics student is free to second-guess them, and it is widely agreed that the judges have often been mistaken. Section 2 cases against established market power offer fascinating tests of one's ability to use economic tools to judge the scope of the market sensibly.

Economies of scale The other key question in a Section 2 monopolization case is whether there are large economies of scale. If the dominant firm can show that it has gained important economies of scale, most judges will acquit it, even if the rest of the case against the monopolist is airtight. Economies of scale are not included in the letter of the antitrust laws, but judges are usually persuaded by them nonetheless.

Of course, whether economies of scale really exist is usually highly debatable from case to case. The defendant will claim that the economies are large, while the plaintiff will usually present experts who say that they are small or nonexistent. Once again, the *logic* of the issue is clear, but the *matters of degree* are uncertain.

Two basic concepts—the extent of the market and of possible economies of scale—are central for understanding the antitrust actions taken toward dominant firms. The agencies try to pick those cases where the degree of monopoly is high, compared to what the economies of scale might require. The agencies also naturally focus their efforts on large firms in major industries, where a single case might yield a large economic gain. Often, too, several private firms have already sued the dominant firm, claiming that it has damaged them. These private suits often stir (or embarrass) the agencies into taking action.

Leading cases Leading cases are summarized in Table 4. Two recent cases show the current issues and define the scope of current precedents.

UNITED STATES V. INTERNATIONAL BUSINESS MACHINES CORPORATION* This was the big IBM case, filed in 1969. After a marathon trial from 1975 to 1981, it was finally dropped by the Reagan administration in January 1982. It had become a mammoth case, with hundreds of lawyers preparing the IBM defense, against about ten lawyers on the Antitrust Division side. The Antitrust Division's side of the case took 726

*These and other antitrust cases can be looked up in two main sources. For past cases, the decisions are in the standard volumes of Supreme Court and appeals court decisions. For current cases, the *Antitrust and Trade Regulation Reporter*, issued biweekly, gives details of the events as they unfold.

Table 4 ***A selection of leading antitrust cases against established dominance***

Case	Origin of Dominance	Year of Decision	Alleged (%)	Market Share Defendant's Version (%)	Final Decision (%)	Result of Case
U.S. v. Standard Oil (oil)	1870s	1911	90	60	90	Convicted. Separated into about a dozen regionally dominant firms.
U.S. v. American Tobacco (cigarettes)	1890s	1911	90	90	90	Convicted. Separated into three firms.
U.S. v. Alcoa (aluminum)	1903	1945	90	30	90	Convicted. War plants sold in 1950 to two new firms (Reynolds and Kaiser).
U.S. v. Du Pont (cellophane)	1920s	1956	70	18	18	Acquitted because market held to be broad.
U.S. v. IBM (computers)	1952–1955	1982*	70	33	—	The case was withdrawn.
FTC-Xerox (copiers)	1961	1975†	90+	~50	—	A compromise, giving access to some Xerox patents.
FTC–Du Pont (titanium dioxide)	1970s	1981	55	42	55	Acquitted. Dominance credited to innovation.
U.S. v. AT&T (telephone equipment and service)	1880s	1982†	100	below 50	—	With conviction likely, the case was settled by compromise.

*The case was withdrawn.
†The case was settled by compromise.

trial days and 104,000 pages of transcript; IBM's lawyers called 856 witnesses and cited 12,280 exhibits.

Yet, the economic question at the heart of the case was simple: Did IBM have too much monopoly power? The suit alleged that IBM had held 60 to 70 percent of the computer market since 1955. Also, the suit alleged, IBM engaged in various anticompetitive acts during the 1960s, such as predatory pricing against certain successful competitors (those actions are discussed in that part of the third main section dealing with pricing tactics). IBM had also gained large monopoly profits on its equity capital—18 percent for more than 25 years.

IBM's defense rested on a market definition that included all equipment used with computers in any way (display units, typewriter terminals, office equipment, etc.). IBM also included most other products made by firms that produce computers. Honeywell, for instance, makes computers plus various other industrial equipment; IBM included all of that other Honeywell equipment in the claimed computer market. In a market defined that

broadly, IBM had a market share of only about 35 percent in the 1960s, well below the traditional 60 percent legal threshold.

IBM also argued that its market position arose from the superior "skill, foresight, and industry" that it had shown in developing computers, rather than from unfair pricing and tactics. At the most, IBM said, it had competed vigorously, but hard competition should not be penalized. IBM claimed to be the most innovative computer company, and to have achieved large economies of scale. Its high profit rate, it said, was merely the financial reward for efficiency and innovation.

Reagan officials accepted IBM's arguments, which fitted with their general policy of letting dominant firms alone unless they could be proved to have committed severe abuses. In dropping the case, the officials argued that conditions in the computer industry had changed sharply since the 1960s.

The case left this informal precedent: ***Dominant firms are free to compete with extreme force, even if they drive other firms from the market. If their actions are not clearly abusive, and if the case lasts so long that the industry can be said to have changed, then the firm will be acquitted.***

THE FTC-XEROX CASE Xerox Corporation attained a complete monopoly of plain-paper photocopying in the United States during 1961–1970. It held crucial patents, developed many others, and practiced extensive price discrimination. The FTC challenged this monopoly in 1972. After private negotiations, the FTC and Xerox reached a compromise settlement in 1975 (moving much faster than the IBM case!). Though the Xerox case was not argued or decided formally in public hearing, the issues were discussed in detail.

The FTC defined the market as *plain-paper* copiers only. There are coated-paper copiers too, but the copies they make are rather different in feel and desirability. Since coated-paper copies are regarded as inferior to plain-paper ones, they would not be substitutable. Therefore, plain-paper copiers were said by the FTC to be "the" market. Xerox had 100 percent of the plain-paper market until 1970, and about 85 percent of it in 1974 (IBM, Eastman Kodak, and others had entered by then).

Xerox's price discrimination showed that it had the intent to eliminate competition by making sharp discriminatory price cuts on products where it faced competition. On products where its market power was higher, Xerox set much higher price-cost ratios. The FTC further claimed that Xerox realized few scale economies, was not an innovative leader after 1965, and made high monopoly rates of return averaging above 27 percent during the 1960s.

In its defense, Xerox defined the market as *all* copying during the 1960s, which would have given it a market share of only 65 percent, and all copying *and reproducing* (mimeographing, etc.) in the 1970s. On that basis, Xerox's market share was probably well below 50 percent by 1973. Xerox also pointed out that the crucial patents it held were perfectly legal and binding. Therefore, it said, its large market share, pricing, and high profits were justified. Xerox also claimed that those profits, which dwindled below a 20 percent return on capital by 1973, merely reflected Xerox's scale economies and fruitful innovations.

You can judge for yourself by direct experience whether plain-paper copies are closely substitutable for coated-paper copies and/or other reproducing processes. Consider their physical features, plus speed and convenience. Most coated-paper copies have had a shiny surface, a gray shading, and a chemical smell. It is debatable whether these make coated-paper copies inferior to plain-paper copies. Would you expect the cross-elasticity be-

tween these methods of reproduction to be high? The claims about scale economies and innovation cannot be proved or disproved, for the facts are unclear and experts' opinions vary.

A moderate compromise was reached in 1975, giving Xerox's rivals the chance to license some of Xerox's extensive collection of patents on copier technology. Little new U.S. competition has emerged since then, but new Japanese products marketed by the Canon and Savin companies since 1975 shrank Xerox's share of plain-paper new-machine sales to less than 50 percent by 1981.

This sample of just two cases cannot show the full variety and lessons from actions against dominant firms. But it does reflect several common features of these cases. The cases can be extremely lengthy, for they are complicated and intensely debated. Indeed, the defendant usually gains by delay and can stall action by many procedural tactics. Even the most basic issues can be made to seem complex during the legal contest, as each side's lawyers strive to win. The details of these cases proliferate, often into scores or hundreds of volumes of testimony, reports, and data.

Under current decisions, dominant firms can expect to continue without challenge, even if their market shares are well above 60 percent. Only if the firms are clearly abusive or inefficient may they be convicted of monopolizing. And even then, the penalties may be light. Since 1913, the courts have rarely required divestiture (that is, the selling off of part of a firm to reduce market share). Severe competition by dominant firms now appears to be largely immune to antitrust policy.

Antitrust policies toward mergers

There are usually at least 1,500 mergers a year in the United States. The antitrust agencies challenge about 30 per year and usually win about two thirds of those cases. The court precedents for merger policy were set mainly by a series of landmark cases in the 1960s.

Leading cases Leading cases are summarized in Table 5. The precedents from these cases probably forestall thousands of other mergers because the potential partners expect that their projected merger would be prevented. We will present four of those landmark cases, to show how merger policies have been formed.

U.S. V. VON'S GROCERY COMPANY (1966): A HORIZONTAL MERGER There were no effective laws against mergers until 1949. After 1910 or so, the dominant firms in perhaps ten major industries knew that any further mergers would probably precipitate a major Section 2 case. But otherwise firms could merge with impunity, even though they could not collude to fix prices.

In 1950, this merger loophole was closed by the Celler-Kefauver amendment to the Clayton Act. There was a pause of six years, until a merger came along to test the law. That merger (between two steel companies) was stopped, and then others involving shoe companies and banks were prevented. Finally, in 1966, the Von's Grocery decision set the strict rules for horizontal mergers. That strict precedent held until 1981, when Reagan administration officials announced a loosening of the rules.

For a case with such a large precedent, Von's Grocery involved remarkably small firms. Two small Los Angeles grocery-store chains had tried to merge in the 1950s: Von's, the third largest chain, and Shopping Bag, the sixth largest. After the merger, their combined share of the Los Angeles grocery market would have been only 7.5 percent. Concentration in the Los Angeles market was declining, and entry barriers in it were low. The merger would

Table 5 ***A selection of leading antitrust cases against mergers and price fixing***

Case	Year of Merger or Price-fixing	Year of Decision	Market Share Held by the Merging or Price-fixing Firms (%)	Action Taken
Horizontal Merger				
Brown Shoe–Kinney Shoe (shoes and shoe retailing)	1957	1962	20	The merger was prevented.
Von's Grocery–Shopping Bag (grocery stores in Los Angeles)	1959	1966	8	The merger was prevented.
Vertical Merger				
Du Pont–General Motors (paints and fabrics)	1920	1957	30	Du Pont's was required to sell its shareholding in General Motors.
Conglomerate Merger				
Procter & Gamble–Clorox (bleach)	1958	1967	55	Clorox was restored as a separate firm.
ITT and various firms (hotels, baking, car rental, insurance, etc.)	1960s	1971	"leading"	ITT chose to retain Hartford Fire Insurance and to sell several other firms.
Price-fixing Cases				
U.S. v. Addyston Pipe & Steel Co. (cast-iron pipe)	1890s	1899	30	Conviction and fines.
U.S. v. Socony-Vacuum (gasoline)	1930s	1940	35	Conviction and fines.
Electrical Equipment Cases (heavy electrical equipment)	1930s–1950s	1960	over 90	No defense: fines and several brief jail terms.
U.S. v. General Electric and Westinghouse (turbine-generators)	1963–1975	1976*	over 90	Compromise. The scheme for tacit collusion was renounced.

*The case was settled by compromise.

therefore not have eliminated much competition.

Yet the Supreme Court stopped the merger. The majority opinion stressed the need to prevent concentration and preserve vigorous small competitors. It was willing to promote low concentration, even down to preventing this small merger. If the Court was wrong, the harm would be slight. Each grocery chain, after all, was still free to grow by setting up new stores or buying them a few at a time.

The Court's minority in this case complained bitterly, saying that since the mar-

ket was already highly competitive, the merger was harmless or even beneficial. Was the decision wrong? Opinions are still divided. Most scholars now think that the Von's Grocery precedent might have been too strict, but not by much.

U.S. V. DU PONT (1957): A VERTICAL "MERGER" The Du Pont chemical company began as an explosives maker. During World War I, it made enormous profits, some of which it used in 1919 to buy about one quarter of the stock of the General Motors Company. In the following decades, General Motors bought most of its paints, glues, and seat cover fabrics from Du Pont. The Antitrust Division sued Du Pont in 1949, claiming that this vertical link between the two companies had excluded other paint and fabric firms from a fair chance to sell their products to General Motors. The Supreme Court agreed in 1957, forcing Du Pont to sell its shares and end the link. Since then, General Motors has broadened its purchases sharply. The case did increase competition in the market.

FTC V. PROCTER & GAMBLE (1967): A CONGLOMERATE MERGER When Procter & Gamble bought Clorox Chemical Company in 1958, P&G was the largest household products firm. It did not sell bleach, but it had been planning to enter the bleach business. Some of its products were related to bleach, and P&G management had considered making a direct entry—by building a new factory to produce bleach—before deciding to enter by buying out the Clorox Company instead. Clorox itself was the dominant bleach firm, with a long-established share of 49 percent of the national market. Clorox's share in the Mid-Atlantic region was as high as 71 percent, compared to 15 percent for Purex, the next largest bleach producer.

The merger would clearly have subtracted a leading new "potential entrant" into the bleach market: P&G itself. That would have reduced competition and, by itself, probably led the FTC to stop the merger. Yet, the FTC (later affirmed by the Supreme Court) instead cited P&G's advertising advantages as the main grounds for preventing the merger.

The FTC and the Court stressed that P&G would be able to give Clorox overwhelming advantages in advertising and distributing its bleach. P&G was the nation's largest advertiser (spending over $175 million on advertising in 1967), and its discounts and market power were likely to entrench Clorox further as the dominant bleach firm. A "toehold" acquisition by P&G of a small bleach company (say, Purex or smaller) would not have encountered this objection and would likely have been allowed. The loss of a potential competitor would have been more than offset by the increase in the small firm's ability to compete vigorously.

ITT's CONGLOMERATE MERGERS During the 1960s, the International Telephone and Telegraph Corporation bought up a series of large firms that were leaders in their markets: Continental Baking ("Wonder Bread," etc.), Avis (car rentals), Levitt (house builders), Sheraton Hotels, Canteen Corporation (dispensing machines), and others. The Antitrust Division sued ITT in 1969, saying that ITT's large financial resources would help to entrench these leading firms even more, thus reducing competition. By contrast, smaller "toehold" mergers—in which ITT bought firms with 10 percent market shares or less—would have promoted competition by building up little firms to compete more effectively with the market leaders.

ITT settled the case in 1972 by selling off some of the firms. Because the case was not tried and brought to a decision, it did not set any clear precedents for other conglomerate mergers. Still, the case empha-

sized that conglomerate mergers with leading firms may be challenged and even stopped.

These cases convey some of the fascinating variety and drama of actions against mergers. Sometimes the agencies go too far, fighting mergers that are neutral or even economically desirable. The courts, too, are fallible, often shifting the policy rules too far one way or the other. The basic issue is always: ***What conditions—in the form of low market shares and low entry barriers—will ensure effective competition and efficient production at an adequate scale of production?***

Where scale economies are large, mergers may merely help to achieve them. But where the economies are small (recall Table 5), most mergers are purely a means to gain monopoly power, for they give little or no gain in efficiency. The economic task is to set the limits on mergers carefully, so that no more competition is sacrificed than scale economies make necessary. Presently, the policies may be roughly correct, setting a fairly strict standard of skepticism toward claims of scale economies.

Even if merger policies err toward being too strict, the harm is usually slight, for firms can always *grow internally*, instead of by mergers. They can build their own new factories, thereby increasing the industry's capacity and competition. The internal growth may take longer and incur some added costs. But it is possible, and it *increases* competition rather than reducing it. So merger policy is relatively easy to apply, because (1) it only stops new conditions, rather than trying to change long-established positions; and (2) it leaves internal growth as a good alternative.

The contrast is sharp between strict merger policy and weak actions toward existing monopoly. It is easy to see why this difference has evolved. Section 2 cases are hard to win, because of (1) complexity, (2) severe resistance and delay, and (3) the reluctance of judges to tamper with large successful firms. Yet there is an awkward gap between the 10 percent ceiling on mergers and the 60 percent safety level for dominant firms. To be economically consistent, the two criteria should be brought closer into line.

Policies toward price fixing and other actions

Price fixing The agencies probably catch little of the price fixing that goes on in oligopoly markets. Even so, the range of cases and convictions is remarkably wide. In a recent six-month period, cases in the biweekly *Antitrust and Trade Regulation Reporter* (which your college library may have) included: Korean wigs, ready-mix concrete, Hawaii package tours, paper labels, timber, Utah egg dealers, steel products, construction firms, bakeries in El Paso, liquid asphalt, plumbing supplies, and scores of others. Even tight oligopolies frequently indulge in elaborate price fixing.

U.S. V. ADDYSTON PIPE AND STEEL COMPANY (1899) In this first landmark case, six producers of cast-iron pipe in the region including Ohio and Pennsylvania had divided up their markets and operated a bidding ring. To prevent competition, they arranged to rotate the contracts among them, designating for each time who would make the lowest bid. Such a bidding ring ensures cooperation among the sellers and gives the buyers no real choice. Though the six firms held less than half of the markets, their price fixing was convicted as illegal *per se* by William H. Taft, then an Ohio judge. Thereupon, the firms soon merged with each other to fix their prices internally, and legally!

THE ELECTRICAL EQUIPMENT CONSPIRACY (1960) Sixty-one years after *Addyston* made price fixing flatly illegal, this spec-

tacular case showed that price fixing had been a way of life for decades in seven major markets for heavy electrical equipment (generators, transformers, switch gear, etc.). Producers of heavy electrical equipment had run secret bidding rings, using formulas based on phases of the moon to rotate orders among themselves. Executives met in hotels, motels, cabins, bars, and other secret retreats. The cloak-and-dagger operations often degenerated into wrangling, but they did raise prices by 20 percent or more for long periods on many billions of dollars of equipment. Some 29 companies, including General Electric and Westinghouse, and scores of their officers, were involved. There were fines and damage suits by customers, and some officials served brief jail sentences. The whole set of penalties, however, was not generally regarded as severe.

THE GENERAL ELECTRIC–WESTINGHOUSE CONSENT DECREE (1976) Even after the electrical equipment manufacturers were caught and penalized in 1960, the industry's tight oligopoly structure remained a basis for tacit collusion. Competition did break out vigorously in 1960–1963. Then, in 1963, General Electric set up a new pricing method, based on simple formulas, which it published in full. Finally GE promised to give any new price cut retroactively to all other purchases made during the *previous* six months. That amounted to a heavy penalty on itself for cutting its prices. Therefore, it was a form of pledge to its rival that it would rarely cut prices.

Moreover, GE pledged to publish all of its price offers and orders.

Thus, GE surrendered all of the secret competitive tactics by which oligopoly pricing can be kept flexible and sharply competitive. This can be illustrated by imagining that GE and Westinghouse have both sold 15 turbines in six months for $20 million each. But GE wishes to get a major new contract for 5 turbines by bidding only $18 million each. The retroactive price cut (to $18 million each on the 15 earlier turbines) costs it $30 million, besides the $10 million on the 5 new turbines. That extra $30 million penalty will discourage GE from making the price cut at all. Moreover, Westinghouse would know exactly what GE would do, so the chances for avoiding competition are high. GE's main rival, Westinghouse, immediately copied GE's plan, down to the precise numbers.

Therefore (as GE memos show was intended from the start), the two companies now had a firm basis for mutual trust and tacit collusion. For many years after 1964, there was no price cutting or flexibility in this industry. Competition had been tacitly eliminated. Only later, after 1970, when a large utility customer (American Electric Power Company) sued GE and Westinghouse, did the Antitrust Division intervene with its own case. Eventually, in 1976, the firms agreed to drop the plan and restore competitive pricing. But there was no trial, conviction, penalties, or payment of damages.

This major industry shows how tight oligopoly poses sharp problems. Tight oligopoly can result in straight price fixing, which often can be caught, proved, and penalized. Even after such treatment, however, the temptations to collude still remain, on a more informal but nearly as effective basis. Changing the industry's structure into loose oligopoly by dividing the leading firms into smaller units usually seems to be too drastic a treatment. But without a structural change, the tendency to collusion will probably remain.

Price discrimination A leading firm can often reduce competition, not by cooperating with its rivals, but instead by taking actions that harm them. Such actions are de-

signed to victimize and exclude competitors. When dominant firms act this way too severely, they can reduce competition.

Some price discrimination does just that, especially when a dominant firm makes deep selective price cuts. Remember that price discrimination can *promote* competition when it is done sporadically by firms with small market shares. Only when it is done systemmatically by dominant firms is discrimination *harmful* to competition. Standard Oil's selective pricing before 1900 often had this effect.

For another example, IBM's new 360 line of computers in the 1960s was threatened at two points: by GE "time-sharing" computers and by computers designed for scientific uses. IBM rushed out two costly stop-gap models to meet that new competition, at prices that did not cover IBM's costs. Such loss-making models succeeded in stopping the competitors, helping drive both GE and RCA from the market altogether. The episode illustrates how a dominant firm can selectively defeat competitors who have superior products by setting prices that are below cost.

There are no recent cases on predatory actions that offer clear precedents and rules. Indeed, judges in the 1970s have generally acquitted "predatory pricing," even when little firms were sharply damaged by the larger firms' tactics. If a dominant firm's selective price cuts don't take price levels below the levels of marginal cost, they will probably be exonerated, regardless of the impact on other firms.

The economic argument for acquittal usually runs as follows: As long as price is not cut below marginal cost, any efficient rival should be able to survive the price cuts. The selective price cuts should be welcomed for providing products to customers at low prices, while weeding out inefficient producers.

On the other hand, even if price does not go below marginal cost, it may go below average cost. The smaller firm would lose money and perhaps face bankruptcy. A dominant firm can usually weather such hard times better than its small rivals. Also, large firms can harm little efficient rivals in many ways, with pricing as only one of their weapons. For example, the large firm can announce a new line of products just when a small firm brings out a new product. That will induce customers to hold off buying the small firm's offering. Or the large firm can merely threaten to cut prices deeply, causing fear and fluctuations for small rivals.

At any rate, marginal cost is usually hard to measure as a basis for judging if the price cuts have been abusive. Thus, a dominant firm's whole strategy can be damaging even if the prices do not go clearly below marginal cost.

Summary

1. U.S. antitrust developed in thousands of fascinating cases since 1890. U.S. antitrust policies are a unique effort to get economic benefits with minimal public cost and interference in industry. Antitrust is an imperfect human activity, as fallible officials and judges process a stream of cases, many of which have confusing details.
2. The two antitrust agencies are small, compared to their economic tasks. The aims of antitrust are both efficiency and equity: to promote competition as far as it is consistent with scale economies, and to promote fair conditions and outcomes.
3. If antitrust works, then competition is more effective and will be more productive, open, and fair. Accordingly, direct control of firms will not be needed.

4. Antitrust policies have been mainly gentle toward existing market dominance, but strict toward mergers and price fixing. This causes some imbalance and unfairness, letting dominant firms do individually what their little rivals cannot do together.
5. Yet, the whole effect of antitrust has been toward much less concentration and price fixing than otherwise would have occurred. The U.S. economy is much closer to the ideals of free competitive markets because of antitrust policies.

Key concepts

Antitrust policy

Questions for review

1. a. What is the major criterion by which public policies should be judged?
   b. Describe the agencies responsible for enforcing antitrust policies.
2. a. Which type of merger is least likely to be challenged (horizontal, vertical, conglomerate)?
   b. What determines whether a merger should be challenged?
3. Price fixing and price discrimination are two offenses which are likely to be treated quite strictly by the courts. True or False? Explain.
4. How can cross-elasticity and economies of scale be used to determine the outcome of antitrust action?
5. Describe some of the potential harms and benefits of conglomerate mergers.

13

Policies Toward Monopoly Power: Regulation and Public Enterprise

As you read and study this chapter, you will learn:

- the economic reasons and criteria for regulation
- four economic issues of regulation: marginal-cost pricing, effects on efficiency, new competition, and deregulation
- economic criteria for public enterprises
- several case studies of actual public firms

When you switch on a light or mail a letter, you are dealing with the subject matter of this chapter. Electricity is a classic case of a natural monopoly, placed under public regulation. The U.S. Postal Service is a public enterprise because it, too, is both a natural monopoly and has certain social purposes.

In these cases and others, competition is not efficient. Instead, the public lets the supplier have a monopoly, and then it regulates its prices. The supplier may also be put under public ownership. These approaches rest on several clear economic concepts, which we present in this chapter. But, in practice, the problems are often complex, and the results are debatable. We present some of those issues too.

In the first section, we discuss the economic regulation of prices. Then, in the next section, we present public enterprise.

Regulation of utilities

In some markets, one or several firms are given an exclusive franchise and then supervised by a regulatory commission. The commission has powers to scrutinize the firm and to control its prices. Such price regulation has covered a series of public utilities and several oligopolies (such as airlines). It is a distinctively American approach, combining a maximum of private ownership with some degree of public control. It is supposed to achieve economies of scale in cases of ***natural monopoly,*** while keeping the monopolist's prices down toward costs.

The economic objective of ***regulation*** is shown in Figure 1, for electricity service, for example. There are large economies of scale, with the average cost of electricity declining to the output level Q_1. The demand curve for electricity intersects the average cost curve at that same output level. The regulators now set the price of electricity at P_1, and so consumers demand—and receive—the output level Q_1. No excess profits are earned by the firm, capacity is fully used, and electricity is supplied at the lowest possible cost. The economies of scale are achieved, while price is held down to the level of cost.

At its best, regulation does apply such controls, briskly and fairly. The economic task has two parts. One is to set price levels, so that the firm does not earn excess profit and exploit its customers. The other part is to set a price structure that is "just and reasonable," among the variety of customers. The monopoly will try instead to set discriminatory prices, along the lines shown in Chapter 10. Economic efficiency requires aligning prices with marginal costs instead. Remember that the alignment of price and marginal cost brings value into line with sacrifice at the margin.

The ideal commission does these two tasks with a minimum of cost and delay.

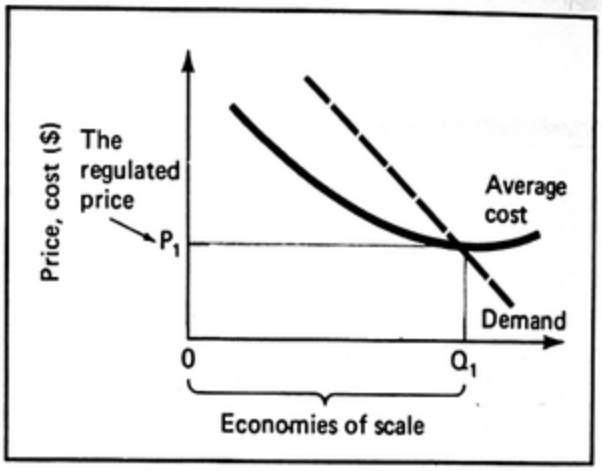


Figure 1 Regulation of a natural monopoly

The large economies of scale are shown by the decline of the average cost curve in the output range of zero to Q_1. The market demand curve passes through the average cost curve at its minimum point at output level Q_1. By setting the maximum permitted price at P_1, the regulators prevent the firm from earning excess profits. Output Q_1 is produced, and the economies of scale are achieved.

And when natural monopoly conditions fade away, then the regulation and the franchise should be withdrawn, so that competition can take over the job.

Yet, regulation may, instead, go wrong. It may become a captive of the industry. It may be applied where natural monopoly conditions do not exist. It may be slow, ineffective, and costly, and it can have inefficient side effects. Regulation now covers industries with only about 6 percent of national income and 15 percent of total investment. Yet, it raises important and complicated issues. And regulation is highly controversial: Is it a charade, as some experts suggest? Sluggish? Highly effective? A captive? A cause of waste? Since 1960, criticism of regulation has been rising, and, in fact, as we noted in Chapter 11, in the 1970s, several major sectors were *deregulated*.

Patterns of regulation

What is to be regulated? Ideally, regulation is applied to natural monopoly, as shown by the down-sloping average cost curve. The firm's resulting monopoly power could be enhanced if (1) the good is a "necessity,"

with highly inelastic demand (such as for electricity, water and telephone service); and (2) users are physically connected to the supplier (as by wires or pipes). In those cases, consumers would be especially vulnerable to exploitation and harmful price discrimination.

These conditions are all matters of degree. Economies of scale are often moderate rather than extreme; industries do not divide neatly into natural competition and natural monopoly boxes. Moreover, technology often changes, so that the econo-mies of scale grow or recede. Today's natural monopoly may soon be naturally competitive. Therefore, the proper scope of regulation is often uncertain and changing, rather than clear.

Commissions There are 4 main federal regulatory commissions and nearly 50 state regulatory bodies. They are summarized in Table 1. There are usually three to seven commissioners, who hear and decide issues brought before them by the regulated firms, by customers, by the commission staff, or by other parties. Commission resources vary from scant to large. Housekeeping and peripheral tasks (such as

Table 1 ***The main federal commissions and five selected state commissions***

Commission (year established)	Number of Members	Number of Staff Members	Budget, 1979 ($ million)	Jurisdiction
Federal				
Interstate Commerce Commission (1888)	11	1,770	29.4	Railroads, trucking, buses, water shipping, oil pipelines, express companies, etc.
Federal Energy Regulatory Commission (1920, 1935)	5	1,191	22.8	Electricity, gas, gas pipelines, oil pipelines, water power sites
Federal Communications Commission (1934)	7	1,785	32.8	Telephones, television, cable television, radio, telegraph, CB radios, ham operators, etc.
Civil Aeronautics Board (1938)	5	708	13.5	Airlines (passenger and cargo), other carriers. (The CAB is scheduled for abolition in 1983.)
State				
California	5	522	12.2	Electricity, gas, telephones, railroads, buses, trucks, airlines, water supply, warehouses, cable TV, sewage, etc.
Colorado	3	72	1.0	Electricity, gas, telephones, railroads, buses, trucks, airlines, oil pipelines, water supply
Georgia	5	57	0.8	Electricity, gas, telephones, railroads, buses, trucks
New York	5	343	12.8	Electricity, gas, telephones, oil pipelines, water supply
Wisconsin	3	140	3.2	Electricity, gas, telephones, railroads, buses, trucks, taxis, airlines, oil pipelines, water, sewage

Source: Federal Energy Regulatory Commission, *Federal and State Commission Jurisdiction and Regulation* (Washington, D.C.: Federal Power Commission, 1980).

safety at railroad crossings and the licensing of small operators) absorb much of the resources of some commissions.

Commissioners are political appointees. Usually they are politically active lawyers, either ambitious young ones or older ones on the way out. Since the more talented commissioners usually rise to higher positions elsewhere, they are in regulatory office less than three years, with little time to develop or change basic policies. Staffs tend to be bureaucratic, cautiously adjusting among the conflicting interests of firms, customers, and other groups. Like antitrust, the process is run by lawyers, who use adversary procedures to turn out decisions meeting legal criteria. The formal legal powers of the commissions are usually large, but the duties and criteria are vague ("fair," "just and reasonable," the "public interest," etc.).

Background The concept of the "independent regulatory commission" was developed in 1885–1910 in the hope of applying expert, honest, nonpolitical control to the problems of natural monopoly. The "utility" firms themselves often lobbied to be put under regulation, since it gave them a monopoly franchise and might be manipulated to serve their own interests. The Interstate Commerce Commission (ICC) was the first federal commission, established in 1888, though it did not gain real powers until after 1910. Wisconsin Progressives started the first state-level commission in 1907. Other state commissions followed, and by the 1930s, most states had regulatory bodies of some sort. The other federal commissions date mainly from the 1930s. Their coverage and activities have evolved with practice and do not fit a uniform pattern.

Until 1944, most commissions were ineffective, stalled by debates over the value of company assets. The firms claimed that the *current* value of assets must be used in setting "fair" profits; but that would have mired regulation in endless, obscure controversies over what the current values really were. In 1944 a landmark Supreme Court decision made the original accounting cost of assets the standard basis for setting profits. This has provided a relatively firm footing for commissions to set strict controls on profits.

A few commissions have applied strict regulation, during some periods. Others have been passive or vigorously procompany. Only in the 1960s did the Federal Energy Regulatory Commission, Federal Communications Commission, and the Civil Aeronautics Board begin to assert firm control over rate levels, rate structures, and the scope of the monopoly held by an individual firm.

Before 1968, there was something of a golden age for most regulated sectors (except railroads). Growth was achieving economies of scale, costs were steady or falling, and the problems to be solved were rather simple. Since 1968, however, severe problems have battered both firms and regulators. These include rapid inflation, ecological impacts, multiplying fuel prices, consumer activism, nuclear power, and antitrust challenges. Regulation has come under great stress, and some commissions have been forced to go deeply into price structure and competitive issues.

Thus far, the 1970s and 1980s have been a watershed, with Congress removing some regulatory controls over airlines, air freight, railroads, trucking, telephones, cable television, banking, and natural gas. The debate and flux continue.

Evolution Most utility sectors evolve through a four-stage process, as shown in Table 2. Stage 1 is the birth of the industry. Stage 2 is rapid growth. Stage 3 brings stability, and the industry matures. Stage 4 is a reversion to natural competition, when regulation is no longer needed. These

utilities are natural monopolies only during the first three stages, when there are large economies of scale. These economies then shrink, which allows competition to exist. Therefore, natural monopoly conditions will usually justify regulation only for a finite period.

Regulation also evolves. It is usually promotional at first, to boost the industry's growth and penetration of the market. Then, in Stages 3 and 4, it often tries to protect the firm from new competition. Deregulating is frequently a difficult process, resisted by the commission *and* by the regulated firms. Therefore, regulation often fits the natural monopoly conditions poorly. Also, the real scope of effective regulation is often different from the area that, by the legal definitions, is supposed to be under control. Even when a commission reaches the right fit, the conditions may soon change and go out of alignment.

Process Commissions hold open hearings on issues put before them and then render decisions. In the typical rate case, the firm announces a new, higher set of prices and asks the commission to approve them. Hearings are scheduled at which the company makes a detailed case for its request, often using expert witnesses as well as company officials. The commission staff then presents a rebuttal, presumably representing the consumers' interests. The staff usually urges setting a lower rate of return and price level, and perhaps a different structure of prices. Other parties may also join in. The hearings often take months, and the ensuing decision may come as much as a year after the original request. The commission usually grants a fraction of the request (half is on the basis of its collective judgment.

The procedures provide due process, with an open forum for all interested parties. Each cites criteria and facts that would favor it. The outcome is usually a compromise among the conflicting interests, stated in terms of some criterion or mix of criteria (fairness, efficiency, etc.).

Decisions on price levels and structures Commissions deal with three main kinds of economic issues: price level, price structure, and the scope of competition.

Table 2 Life-cycle stages of typical utilities

	Birth of the Industry	Rapid Growth	Maturity	Reversion to Natural Competition
	Stage 1	Stage 2	Stage 3	Stage 4
Manufactured gas	1800–1820	1820–1880	1880–1920	1920–1950
Natural gas	1900–1930	1930–1950	1950–	
Telegraph	1840–1850	1850–1916	1916–1930	1930–
Railways				
All	1820–1835	1835–1910		
Passenger			1910–1935	1935–
Freight			1910–1960	1960–
Electricity	1870–1885	1885–1960	1960–	
Street railways	1870–1885	1885–1912	1912–1922	1922–
Telephone	1875–1880	1880–1947	1947–	
Airlines	1920–1925	1925–1965	1965–	
Television	1935–1947	1947–1965	1965–	
Cable TV	1950–1955	1955–		

Source: W. G. Shepherd, *The Treatment of Market Power* (New York: Columbia University Press, 1975), p. 228.

Price level is the conventional topic, refined by decades of practice to a traditional litany of issues. The elements are summed up in the following equation:

$$\text{Permitted rate of return} = \frac{\text{Profits}}{\text{Invested capital}}$$

$$= \frac{\text{Total revenue} - \text{Total cost}}{\text{Invested capital}}$$

The commission decides what the firm's "rate base" is (its amount of capital invested in the business). Next, it decides what rate of return is "fair," usually in the range of 7–11 percent. Then the firm is allowed to set price levels that will generate enough sales revenue to provide the fair rate of profit on the rate base. Hence, this approach is often called "rate-base regulation."

If the commission permits higher output prices, that will raise total revenue, increase profits, and therefore raise the rate of return. The company wants maximum profits, while the commission tries to hold profits (and prices) down to much lower levels.

The basic choice is shown in Figure 2. The utility firm is assumed to have built the right level of capacity; the demand curve cuts both the average and marginal cost curves as close as possible to the minimum of average cost. Therefore, ***a price set at marginal cost will give economic efficiency.*** **Marginal-cost pricing** will also avoid excess profits and will achieve the lowest possible average cost.

The utility would prefer a higher price. To maximize its profits, it would like to set output at Q_m, where its marginal cost just equals its marginal revenue. The price would then be P_m, and profits would be maximized and high (note the large Excess profit rectangle in Figure 2). Since the regulatory commission instead tries to limit price to P_c, there is inherent conflict between the regulators and the firm.

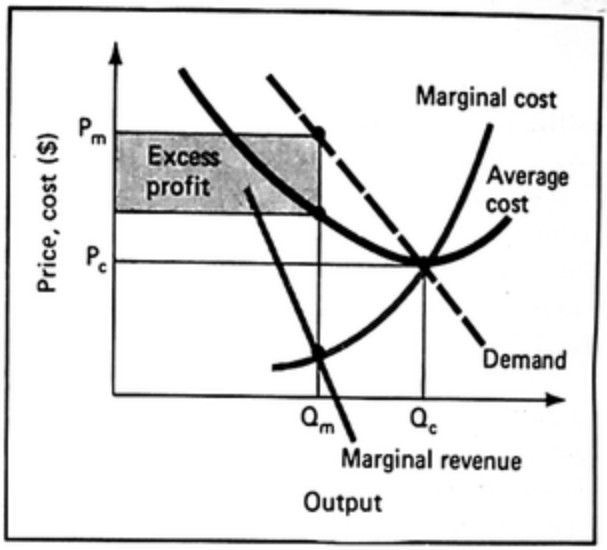


Figure 2 The basic economics of utility regulation

Capacity (the bottom of the average cost curve) is roughly in line with demand. The utility firm would like to set output at Q_m and charge P_m, making excess profit, as shown. But the regulators try to set price at P_c, which gives the utility enough total revenue to cover its total costs. The utility is required to produce Q_c, which is the amount that people want to buy at the regulated price P_c.

In practice, the commissions' decisions usually have no clear rationale. Yet, beneath the process lie some remarkable economic issues.

Some ceiling or "permitted" rate of return is to be set by the commission, but its level is controversial. The laws usually require a fair rate of return, neither too high (unfair to customers) nor too low (unfair to the firm's shareholders). It should also be efficient, by several possible criteria: It should equal the cost to the firm of its capital (the "cost-of-capital" criterion). And/or it should be high enough to attract just the optimal amount of new investment (the "capital attraction" criterion). And/or it should be in line with the risk-return conditions in other industries (the "comparable returns" criterion).

These three criteria all relate to the same basic concept of efficient allocation of capital. But they are not precise as guides to real conditions. "Fair" rates of return usually lie between 6 and 12 percent, but the correct level for each case can be debated endlessly without arriving at a

definitive answer. The commission simply applies its judgment and picks a figure or range, such as 10.25 percent or 9.5–11.0 percent.

Then the value of the **rate base** is fixed by the commission. The firm's invested capital includes (1) fixed capital, at various possible depreciation rates; and (2) other assets, including a range of short-term and liquid assets. Some or all of this is allowed in the rate base, in what can be a complicated judgment by the commission.

Total costs may also be reviewed, to make sure that they are necessary and not inflated—in our terms, to assure that they are "X-efficient." The specific price level then follows fairly directly, since it is the price change needed to let the firm's profit rate go up to the permitted ceiling rate.

These price decisions usually ignore two complications. First, demand may be elastic. Since price changes will alter the amounts consumed, the net revenue change may not be a simple matter at all. Second, future conditions may change, so that the new price schedule turns out to yield profits either above or below the permitted rate of return. Indeed, actual profit rates often do rise above the permitted ceilings.

The decisions are usually only a prediction about the price level that will actually result in the optimal or reasonable profit. Moreover, beneath the veneer of arcane debates about criteria, they are usually just a compromise. Such rough-and-ready decisions must be made, and the regulatory outcomes may even turn out to be reasonably close to the ideal solutions.

Price structure is supposed to be "just" and "reasonable," in the standard legal wording. Price discrimination by these firms is likely to be very sharp; they have a complete monopoly, and they sell to a wide variety of customers (in homes, in shops and factories of all sizes) who usually have very different demand elasticities. Some degree of discrimination may be efficient; but that is a very complex issue, beyond the scope of this chapter. Generally, optimal pricing would contain much less price discrimination than the firm would prefer.

Instead, the proper criterion for prices is cost—specifically, marginal cost. For each specific customer group, price should be set as close to marginal cost as possible. That will bring the utility into line with efficient allocation in the rest of the economy.

The structure of costs may be quite complicated. The regulatory task is to bring prices at least roughly into line with that cost structure, while avoiding discriminatory patterns. Overhead and joint costs (costs incurred supplying all customers) often make marginal costs unclear. Also, most regulated utilities have marked fluctuations in demand, such as the peak loads for electric and telephone service during business hours, and off-peak levels during nights and weekends. These fluctuations cause marginal costs to vary sharply, being high at peak times and low at off-peak times. Therefore, the efficient price structure will also need to have marked differences—by seasons, by day, and by time of day, as we will analyze shortly. The topic can grow difficult, obscure, and frustrating in actual hearings.

Until recently, most commissions have allowed firms to decide most of their price structures. The firms, in turn, have tended toward (1) discrimination or (2) flat across-the-board price changes that, being uniform, minimize complaints among customer groups. Since about 1965, price structure has received closer attention from some commissions.

Four economic issues of regulation

As noted, the four main economic issues of regulation are: setting prices in line with

marginal costs, the inefficiencies that regulation may cause, competitive pricing, and deregulation.

Marginal-cost pricing Regulated firms usually have a variety of outputs, differing by physical features (size, weight, design) or by conditions of supply. Seemingly uniform products can differ sharply in costs. For example, the cost of a kilowatt-hour of electricity at midnight will differ from that of one at noon. More generally, off-peak production usually is cheaper than production at peak-load times. That is shown by the typical daily load curve in Panel I of Figure 3. Output peaks during the day, and then falls to low levels during the off-peak night-time hours. The best equipment is run continuously, giving low costs at off-peak times. That corresponds to low marginal cost in Panel II. But at peak times, costly extra capacity must be started up and used, at high marginal costs. Therefore, peak-load marginal costs are commonly a multiple of off-peak costs.

But remember that *price should equal marginal cost.* If price diverges sharply from marginal cost, then allocation is inefficient. Therefore, utility regulators should strive to get utility price structures into line with marginal costs. That calls for *peak-load pricing*, with prices set much higher at peak times than at off-peak times. In Panel II of Figure 3, the efficient prices are P_o and P_p, with outputs at *B* and *C*. A single uniform price P_u would in-

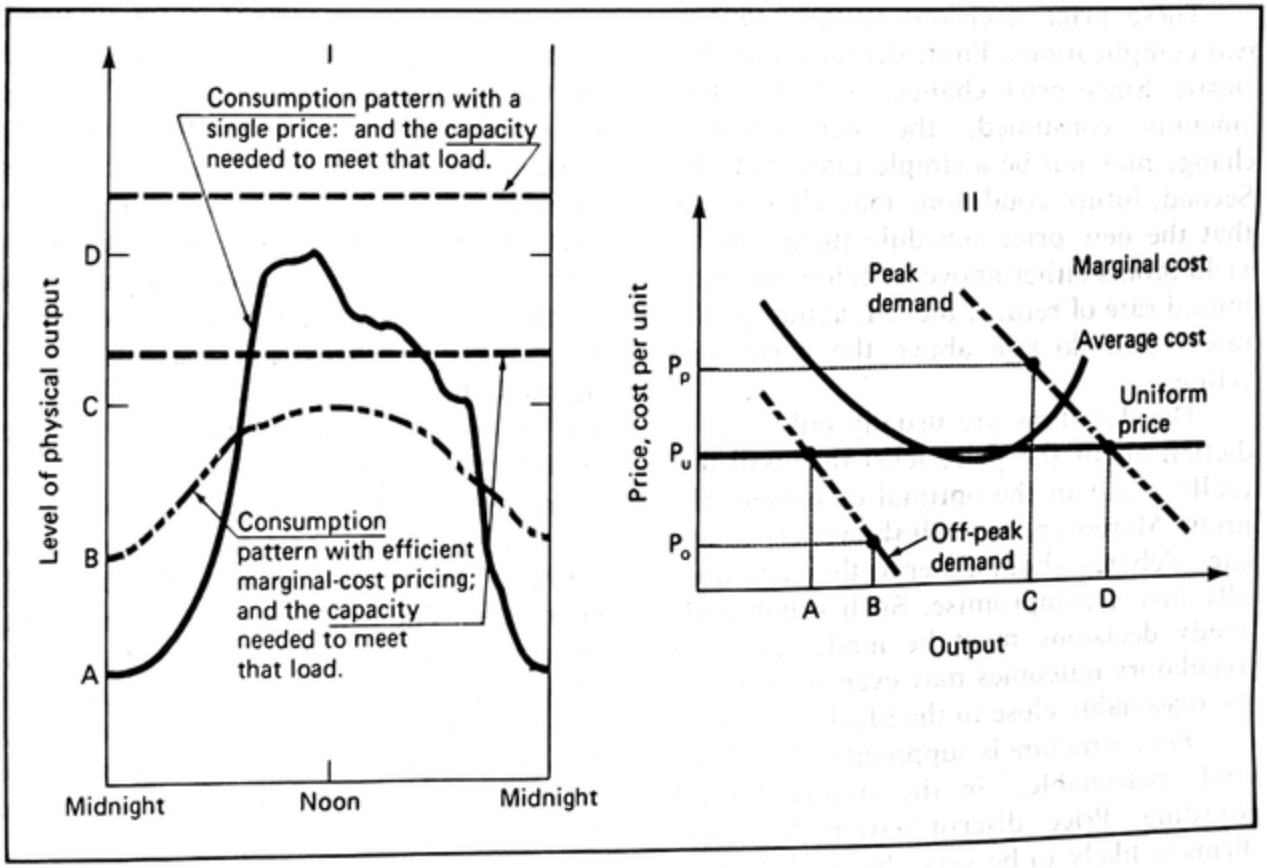


Figure 3 Load patterns, demand, and cost in utility pricing

The letters *A* through *D* are aligned between the two diagrams. The load fluctuates sharply between Levels *A* and *D* if a uniform price is charged at all times. That uniform price is P_u in Panel II. It results in the black-line load curve in Panel I. But if prices are set equal to marginal cost during the peak and off-peak times (P_p and P_o), then the load pattern will be smoothed: it is reduced to *C* at peak times and raised to *B* at off-peak times, as shown by the blue-line load curve.

stead lead to too much quantity demanded at D, while cutting off-peak outputs to Level A.

Such marginal-cost pricing is socially efficient, and often the regulated firm would gain greatly from adopting it. For example, setting prices too low for peak outputs could encourage too much load on the system at peak times and threaten the whole utility system with collapse. Indeed, marginal-cost pricing often lies in the same direction as price discrimination, at least for parts of the utility's output. For example, low-cost bulk power may go to large users who have high elasticity of demand. Both cost and demand would then call for a low price.

Yet, cost and demand conditions often diverge, so that the regulators must force the firm to follow efficient, marginal-cost pricing. Setting high prices at peak-load times is especially important. But it is often hard to enforce because it usually requires higher prices for the periods when the system seems to be "most urgently needed." Also, rate-base regulation may encourage the regulated firm to add more capacity than is efficient. In the great mass of regulated outputs, marginal-cost pricing is both correct and reasonably practical to accomplish.

Nevertheless, before 1965, these lessons were largely ignored, for utilities were eager to raise their growth by means of promotional pricing, which is often discriminatory. Peak-load output was usually priced low, at average costs or even at zero (for local telephone calls, for example). The new scarcities and stresses that have arisen since 1965 have made marginal-cost pricing seem wise, even urgent, both to many regulators and to the firms themselves.

Electricity prices used to ignore high peak-load costs almost entirely, thereby encouraging people to use electricity out to Level D of Figure 3. This, in turn, required the companies to build too much capacity to meet those overstimulated peak-load levels. Local telephone pricing has been even worse. By charging a *zero* price for local calls, the firms have encouraged too high a level of use. That is shown more precisely in Figure 4, where calls at level Q_1 have a true marginal cost that is very high. Thus, when you call a friend at 4 p.m. and chat for a half hour, the true cost may be a dollar or more, although the call's value is low, as shown by the demand curve. But since you pay nothing extra, you (and millions of others) make the calls, and the system absorbs those extra costs.

Long-distance prices reflect marginal costs more closely than do local-service prices. The price differences are familiar and can be seen in the front pages of any telephone directory. Prices during business hours are often double or triple the off-peak (night-time and weekend) prices. Therefore, marginal-cost pricing has been routine in certain utility services, even though it has been avoided in most others.

In the 1970s, there were new efforts to shift toward marginal-cost pricing. In electricity, perhaps a third of the companies have adopted time-of-day pricing. A typical time-of-day price schedule is shown in Table 3 for Wisconsin Electric Power Company. Peak times are defined rather roughly as 7 a.m. to 7 p.m. Monday to Friday during the seasonal hot-weather peak in July–October, when the heavy use of air conditioners strains the whole electric system's capacities and makes marginal cost very high. Off-peak hours are 7 p.m. to 7 a.m. and weekends. During November–June, the 7 a.m. to 7 p.m. hours are an in-between category, neither peak nor off-peak.

Notice that peak-load prices of 8.2 cents per kilowatt-hour are set far above the off-peak price of 1.3 cents per kilowatt-hour. The in-between period has an inter-

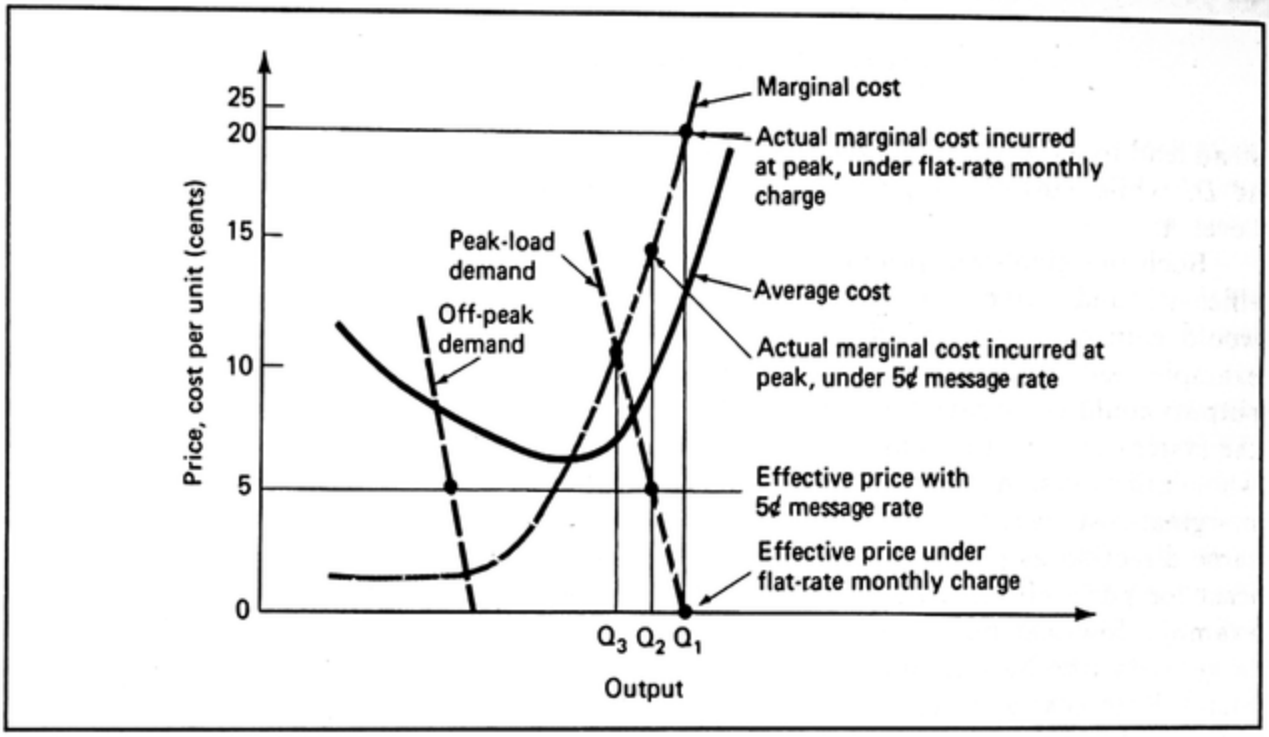


Figure 4 An illustration of local-service telephone pricing

Users pay a flat monthly fee (such as $11 per month) and get unlimited free calls. At that zero price per call, they use service out to level Q_1. At Q_1, the marginal cost of service is high, and the inefficiency is shown by the gap between price and marginal cost.

The efficient peak-load price is where demand intersects marginal cost, giving output level Q_2. A "message rate" may be set, such as 5 cents per call. But it still ignores peak–off-peak cost differences. It would give the output Q_2, with still a large gap between price and marginal cost.

Table 3 A time-of-day price structure for electricity (Wisconsin Electric Power Co.)

Class of Service	**Residential–Time-of-Use**	
Effective in	All Areas Served	
AVAILABILITY		
To residential customers contracting for electric service for domestic purposes for a period of one year or more.		
RATE		
Customer Charge, including one meter	$5.00 per month Billing periods	
Energy Charge per kWh	July–October	November–June
On-peak energy*	8.20¢	5.20¢
Off-peak energy†	1.30	1.30
Meter Charge		
The monthly meter charge for each meter in excess of one shall be $2.50.		

*Residential on-peak energy usage is the energy in kilowatt-hours delivered between 7:00 a.m. and 7:00 p.m. Central Standard Time, Monday through Friday, including holidays.
†Residential off-peak energy usage is the energy in kilowatt-hours delivered during all hours other than during on-peak hours.

mediate price—5.2 cents per kilowatt-hour. These prices fit the cost patterns shown in Figure 3. Even if they are not fine-tuned to fit marginal cost closely, these peak-load prices do at least fit the main patterns of cost. Therefore, they give a practical instance of efficient pricing under regulation.

Despite some progress in this direction, much pricing of electricity, gas, and telephone service is in the old uniform-price patterns that ignore marginal costs. Many economists, therefore, continue to criticize those policies. They have also studied the cost and demand conditions intensively and have developed detailed proposals for revised prices.

The effects of regulation on costs Standard regulation lets the firm charge prices that will cover its costs plus a "fair" profit. This "cost-plus" approach may permit or even encourage X-inefficiency in the firm. If its monopoly power is sufficient, the firm can raise prices enough to cover its costs even if they are greatly inflated. This is reinforced by the firm's interest in providing high-quality, reliable service, which usually entails extra costs. It is difficult to set the socially efficient level of quality and reliability, and the cost-plus-profit basis of regulation may induce the firm to choose too high a level of quality and cost.

In the total cost part of the basic regulatory equation, both the prices and the amounts of the inputs may be raised because of regulation. That is because the firm gets the profits whether it keeps input costs down or not. This problem of "cost-plus" inefficiency has long been familiar in military weapons buying by governments. Under regulation, it is more subtle but still chronic. There are two main limits on it: (1) the professional standards of the industry (managers and engineers presumably apply good sense and technical criteria to what is needed in their system); and (2) scrutiny by the commission (from the start, regulators and courts have recognized the need to guard against possibly extravagant or unnecessary costs). In practice, the controls have usually been weak. The firm's expenses are often listed and looked over in some detail, but little can be done to challenge or rectify dubious cases.

Investment may also be too large under regulation. The conventional method of rate-base regulation encourages the firm to increase the value of the rate base itself. Normally, the permitted rate of return is set a little above the cost of capital. The firm's shareholders, therefore, gain a little (or a lot) of profit from each extra bit of capital included in the rate base.* The process probably works subconsciously, but it encourages the firm to use more capital than is economically efficient.

The rise could come about in two ways: (1) Actual investment could be higher. In choosing new technology, the firm would lean toward more capital-intensive methods. Capacity to meet peak loads might also be higher because of the rate-base effect. This would give the firm more security from embarrassing breakdowns at peak times. (2) Accounting choices would be made so as to maximize the recorded value of assets. Depreciation methods would be the main item to be adjusted, toward writing down the assets' value slowly. The firm may permit overcharging in the prices of the equipment it buys.

The whole rate-base effect has never been accurately measured, and, of course,

*Suppose that the cost of capital is 7 percent and the permitted rate of return is 9 percent. Then every additional $100 million in the rate base will increase net profits by $2 million (that is, $9 million return minus $7 million cost of capital). Capitalized at a 10:1 ratio, that $2 million might equal $20 million in added stock value to the shareowners.

the firms deny that it occurs at all. It probably does shift the margin of choice by some degree in most regulated utilities.

Cream skimming and competition All utility industries have some markets that can be supplied competitively. The regulated firm, however, naturally wishes to encompass them in its exclusive franchise. Indeed, the rate-base effect encourages it; the firm wants to add to its rate base the capital in the adjacent market. Meanwhile, other firms want to get in to compete against the utility.

The key point is that the newcomers are often naturally attracted to the most lucrative parts of the regulated firm's market, where the price-cost ratios are highest. Since there is price discrimination, the entrants usually fasten first onto the "creamy" markets. This "cream skimming" (the British call it "picking the eyes out of the market") is regarded as an acute threat by the regulated firm. The firm will claim that the cream skimming strikes at the "system integrity" of the utility, for the creamy parts are necessary to support the skim parts. With the cream gone, either (1) the whole system will go bankrupt, or at least, (2) prices for most consumers will have to rise.

The regulated firm, therefore, resists any and all competition. If competition is permitted nonetheless, the original firm demands the right to meet the competition by selective price cutting. But if that is permitted, the price cuts may be deep and predatory enough to keep out competition, while still maintaining a discriminatory price structure. The commission thus gets drawn into setting *floors* on specific prices as well as *ceilings* on the firm's whole price and profit levels. And it must usually rely on cost figures prepared by the regulated firm itself.

This baffling problem stems from the basic sources of natural monopoly: overhead costs and economies of scale. These conditions make the regulation necessary, and yet the natural monopoly basis does not extend throughout the system. Competition can, and probably should, enter into some parts. But which parts, how far, and on what competitive terms, must be decided somehow by the commission. Often the conditions are highly complex and changing, and the pressures are intense. Moreover, since commissions are usually imperfect and short of resources, their treatments, too, are often inefficient. They may give the utility firm too wide a franchise or be slow to let in new competition.

The difficulties are widespread in postal service, airlines, railroads, telephones, banking, electricity—anywhere that a commission has to supervise a firm with an exclusive franchise. Cost and competitive conditions vary by gradations, rarely fitting into neat boxes. Regulators are forced to cope with these severe problems as best they can.

Deregulation In extreme cases, regulation should be withdrawn entirely. Such ***deregulation*** often looks attractive, compared with the regulatory effects we have just reviewed. But it must be done with care and sophistication. Since 1975, deregulation has been extensive in parts of the transport, broadcasting, and telephone sectors.

The hardest task is to balance between (1) letting new competition in and (2) withdrawing controls on prices. If the controls are removed before competition is effective, the utility firm has a bonanza: It holds monopoly power but is constrained neither by competition nor by regulation. Naturally, the utilities call for such a "freedom to compete," even though it would be premature and lead to the usual social

costs of monopoly. The opposite error is to let competition in but keep rigid controls on the original firm. Then the firm may be limited too tightly.

Airlines provide the leading case study of deregulation. The economies of scale in airline service are moderate, permitting effective competition on most of the hundreds of airline routes connecting pairs of cities. Yet from its creation in 1938, the Civil Aeronautics Board (CAB) protected the market positions of the 12 original airlines. New airlines were not permitted to enter, nor, with rare exceptions, were existing airlines permitted to move into new routes. Moreover, the CAB permitted the airlines to set ticket prices and then enforced those prices against any competitive price cutting.

During 1960–1975, a series of economic studies showed that these rigid policies were causing inefficiency. A shift to open competition would reduce ticket prices, increase the variety of choice, and improve the efficiency of scheduling flights. The CAB reversed course in 1975 and began permitting new entry and flexible pricing. In 1978, a law was passed to abolish the CAB by 1983.

Airline competition quickly became intense, bringing precisely the improvements predicted by economists. Two possible disadvantages have turned out to be mild. One is that the large airlines have withdrawn from many of the small-city routes, because the sparse traffic does not fill their large aircraft. Yet, small commuter airlines have sprung up to help fill that gap. The other is the heavy use of special discount fares (e.g., "super-savers") that involve price discrimination and could be anticompetitive. Yet, in fact, most of the discounting is procompetitive because the airlines now usually lack dominant market shares.

This remarkable deregulation process has been managed well by the CAB, which has relaxed regulatory limits in balance with the rise of competition. Deregulation has also begun in such other sectors as telephone service, railroads, trucking, and broadcasting.

Public enterprise

A public enterprise *is owned by the state, on behalf of the citizenry.* It can be identical to private firms in every respect, other than having private stockholders. It uses inputs to produce outputs, and it keeps thorough accounts of costs, revenues, and profits. But despite these parallels, the public firm need not maximize its profits as a private firm does. It may pursue other goals, and so its economic performance may differ sharply from that of a private firm. Therein lies the fascination of public enterprise, for it offers a wide variety of possibilities and outcomes.

Coverage and purposes

The main lines of public enterprise in the United States and Western Europe are indicated in Figure 5. The United States differs from other Western economies chiefly in the low share of public enterprise in its utilities, industry, and finance. Otherwise U.S. patterns are not peculiar. The typical pattern in Western economies is (1) *utilities*, entirely or mainly publicly owned; (2) *finance*, one or several public banks; (3) *insurance*, large social insurance programs; (4) *industry*, several major industries under partial public ownership; (5) *social services*, mainly under public ownership; and (6) *distribution*, with little public enterprise.

Public enterprise exists in many parts of the U.S. economy. There is a great variety of forms and behavior, as suggested by

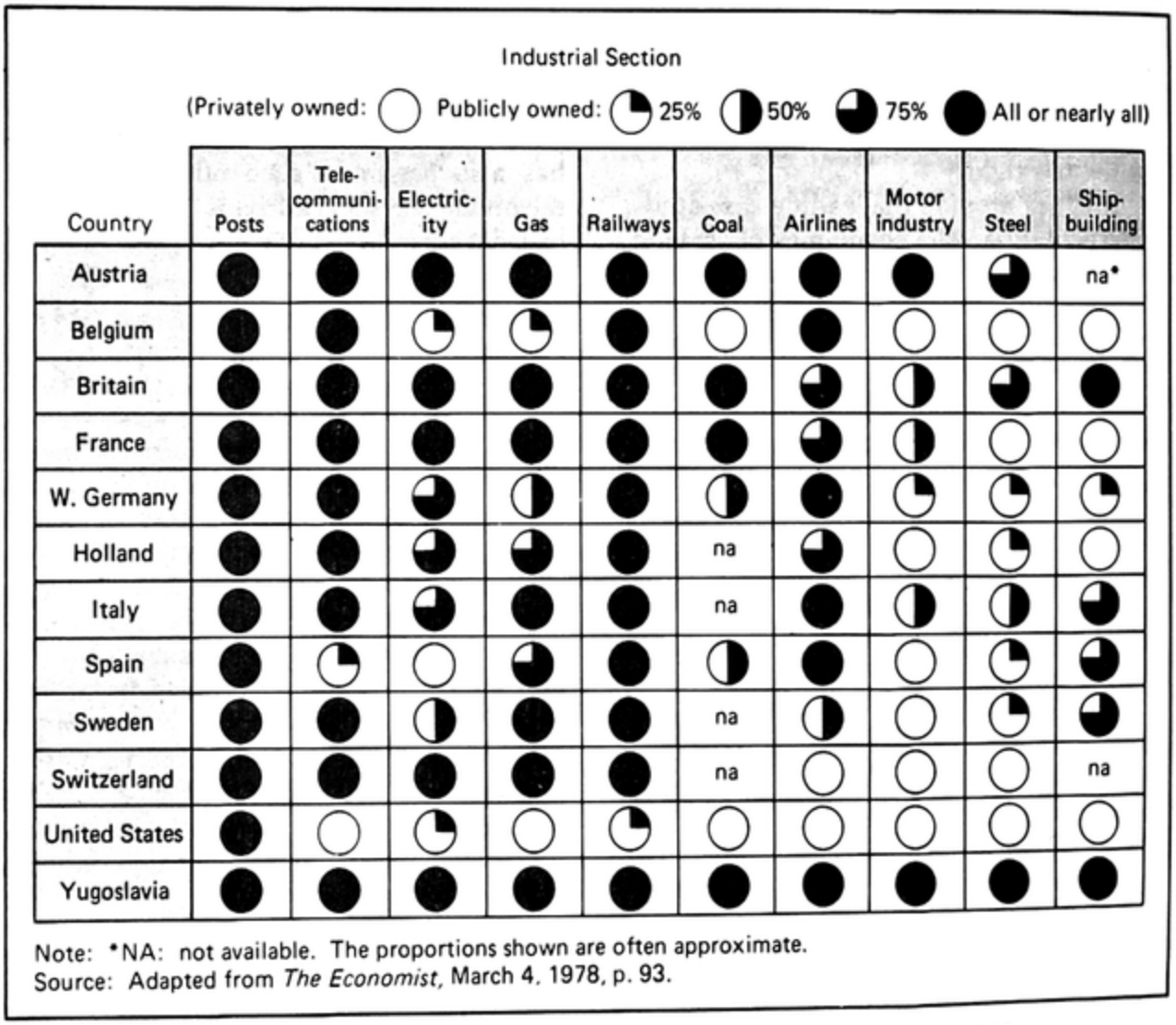


Figure 5 The share of public ownership in major sectors in selected countries, 1978

Table 4. They range from conventional utility cases, such as the Tennessee Valley Authority, to industrial and service areas, over into certain subsidy programs, and into important *social* enterprises such as public schools and universities, mental hospitals, the courts, and prisons. Yet, these public enterprises tend to be a phantom presence in the United States, not recognized for what they really are.

There are many reasons for creating public firms, but the most valid normative reason is that the enterprise can serve some social purpose that a private firm would ignore or violate. This social purpose usually falls under the following headings:

1. *Social Preference.* A society (city, state, or country) may simply prefer public to private control, especially for certain prominent sectors. Such cultural preferences seem to explain much of the great variation in Figure 5.

2. *Inadequate Private Supply.* A new industry or project may seem too large and risky for private firms to invest in. They will demand government guarantees, grants, or other subsidies. It may seem wiser to put the unit under direct public ownership.

3. *Salvaging Firms.* The public often "rescues" failing firms by buying out their

Table 4 ***Local, state, and federal public enterprises in the United States***

	Extent of Public Enterprise
1. Localities	
Utilities	
Transit (bus, subway, trolley commuter lines)	All large cities
Water and sewage	Virtually all large cities
Garbage disposal	Most cities
Electricity	Over 1,000 smaller cities; several large cities, including Los Angeles
Ports	Port of New York Authority (transport and urban facilities); New Orleans, ocean ports
Airports	All large cities
Social Units	
Schools	All cities and towns
Libraries	Virtually all cities and towns
Parks, golf courses, pools	Virtually all cities
Sports stadiums	Many cities
Museums	Many cities
Zoos	Several large cities
Cemeteries	Most cities and towns
2. States	
Prison facilities	All states
Insurance services	Unemployment: all states Workman's Compensation: 18 states
Parks	Most states
Liquor retailing	16 states
Electricity	All Nebraska, a large share of New York
Toll roads, bridges, and tunnels	29 states
Health care	Mental and old-age institutions
3. Federal	*(Expenditures)*
Electricity	Corps of Engineers, $1,420 million; Bureau of Reclamation, $618 million; Tennessee Valley Authority, others
Postal service	$17,700 million expenditures; $784 million subsidy
Lands	Forest Service: $834 million National Park Service: $364 million
Commodities stockpiles	Value about $700 million
Transport	Alaska and Panama Canal railroads; military air and sea transport services; St. Lawrence Seaway
Loans and guarantees	About 100 agencies, includes housing, farming, rural electricity and telephones, Export-Import Bank, Small Business Administration
Insurance	Many agencies: banks, housing, crops, shipping, foreign investment, stock markets, veterans life and annuity insurance, old-age pensions
Health care	Medicare, Medicaid, veterans' hospitals
Industry	Various: Government Printing Office; military production, etc.

Sources: For figures, U.S. Government, *Budget* (Washington, D.C.: U.S. Government Printing Office), Appendix volume.

capital and supporting their rehabilitation. There are always new candidates for such salvage operations. Some are valid. But they tend to burden the public with sick industries that absorb large subsidies.

4. *External Impacts.* Public firms may allow for outside social harms or benefits that private firms ignore. In the extreme, the service may be a *pure public good* calling for a full subsidy (see Chapter 17).

5. *Sovereignty.* A country may take over the local branches of large international firms in order to neutralize their power.

The typical public firm, therefore, has a ***social element*** *to serve, which is apart from its usual commercial goals of producing its services efficiently and selling them at prices that fit cost and demand conditions.* For example, a local bus line is supposed to provide reliable service throughout the city, on a more extensive schedule than a strictly commercial bus line would provide.

The social element is usually debated intensely, both its nature and its extent. What social element is provided by the Postal Service, for instance? And does it require daily deliveries, including Saturday? Should "junk mail" be subsidized? If so, to what extent? You may have noticed the ongoing controversies over Amtrak's services, library hours, parks, Medicare, and city sports stadiums used by professional teams. Quieter debates continue constantly about city services, public schools and universities, airports, golf courses, state liquor stores in 16 states, and all other public enterprises. In every case, the questions are: What is the valid social element? How much of it should the public pay for?

Subsidies and efficiency

The public pays by means of subsidies, which come from government tax revenues. The subsidy can be of any amount, ranging from 100 percent to zero. Thus, the public schools are subsidized totally from taxes, while local water supply is paid for by the users. Most public universities are in between, supported partly by government subsidies and partly by students' tuition payments.

The subsidy ought to be fitted precisely to the social element of the public firm. ***A small social effect requires little or no subsidy, while a large social element might justify a total subsidy.*** Total subsidy means that the direct users pay nothing; the taxpayers pay for it all.

There are two dangers from subsidies to public firms. One is that the subsidy will simply be too large, giving the users an undeserved free ride. Should library users, or local golfers on the public course, or bus riders, or students at public universities be subsidized heavily? Does the service meet a special social need? Are the users really needier than the cross section of tax payers?

The second risk from subsidies is that they will weaken the enterprise's incentives to cut costs. Whenever costs can be covered without effort, the firm may let them rise. The subsidy can become a self-creating device. Public firms as diverse as city transit, the Postal Service, and Medicare are regularly accused of such wasteful and demoralizing subsidies.

These dangers are real, and they have no universal solution. Society must struggle along with its public enterprises, trying to fit the subsidies to the true social element and trying to avoid wasteful incentives. If the political process works well, it may supervise the firms effectively and trim their subsidies to just the right patterns. Public enterprises can go beyond the

narrow limits of profit to serve genuine public needs. But this capacity needs constant control to keep the firms from wasteful mistakes.

Efficient pricing Public enterprises come under the same rules for efficient pricing that private firms do. Their prices should be aligned with their marginal costs (including social costs), just as for regulated utilities. Many public firms do, in fact, adopt efficient price structures, carefully measuring marginal costs and setting prices in line with them. The task is easier because the firms are not subject to the special biases—from monopoly power and cost-plus-profit regulation—that privately owned utilities have.

Yet, many public firms set inefficient prices, and governments often fail to press the firms to improve their policies.

Summary

1. Regulation is a unique U.S. policy. It attempts to limit private firms to zero excess profits and to efficient price structures. There are economic guidelines for these decisions, but the commissions often have to make rough decisions and compromises.
2. Marginal-cost pricing is usually the correct guide for efficient pricing, but it often conflicts with the utility's preferences. Peak-load pricing is being increasingly applied in electricity and telephones.
3. Regulation may induce various kinds of inefficiencies. It may also need to be removed as the sector evolves back toward natural competition. But that transition requires a delicate balance between competition and control.
4. Public enterprises commonly have a social element, as well as commercial operations. Any public subsidy needs to be fitted to this social element. The danger is that the subsidy will diverge from that level, and that it will sap the firm's incentives for efficiency.
5. The actual performance of public enterprises ranges from excellent to poor. Good performance usually requires careful supervision and clear economic guidance.

Key concepts

Natural monopoly
Regulation
Marginal-cost pricing
Deregulation
Public enterprise
Social element

Questions for review

1. a. What is meant by *rate-base regulation?*
   b. What are some of the difficulties inherent in setting prices through rate-base regulation?
2. The price of phone calls usually varies with the time of the call. Explain how this price variation could encourage an efficient allocation of resources.
3. First-class mail users pay 65% of all postal revenues, although lower-class mail (advertising, newspapers, magazines) weighs far more. Do first-class mailers therefore subsidize bulk mailers? Explain.

4. What is the extent of public enterprise in the United States? How does it compare with foreign countries?

5. Are public firms in the United States justified? Explain and give some examples to support your answer.

14

Input Markets

As you read and study this chapter, you will learn:

- the firm's precise choices in buying inputs
- conditions governing the demand for and supply of inputs
- the effect of monopoly upon input choices
- the causes and meaning of economic rent

You have probably seen schematic drawings of the human circulatory sytem, with its miles of large and small blood vessels. One half of the system is arteries, through which fresh blood is pumped to tissues in all parts of the body. The other half is veins, which bring the used blood back to the heart for further circulation. These two complex sets of blood vessels coexist in the complete system.

In the same way, the economic system contains two complicated sets of markets—input markets and output markets—both of which are necessary to complete the system. Output markets were explained in Chapters 4–13. Now we present the other great half of the system, input markets. They too are important, for the factors of production must be priced and chosen for economic activity to occur. They determine such conditions as: wages, which range from less than $3 per hour for some workers to over $350 per hour for others; decisions affecting $5 trillion of capital; and land prices ranging from $1 to $1 million per acre.

Labor, capital, and land are priced and hired in numberless input markets every day. We explain that process in this chapter. Later chapters consider the individual factors more fully: labor in Chapter 15, capital in Chapter 16, and natural resources in Chapter 21.

We begin in the first main section with the individual firm's demand for inputs. By deciding how to use each input, the firm is completing its whole set of profit-maximizing decisions. The second main section of this chapter treats the supply of inputs and the market-wide outcomes, and also presents the concept of economic rent.

The demand for inputs

To present input pricing clearly, it helps to begin with purely competitive factor markets. The first task is to explain how much of an input is used at each price of the input.

The level of input use

The critical assumptions are three:

1. The firm is a profit maximizer in its decisions about inputs as well as outputs.
2. The firm is also a price taker in all *input* markets, where it buys its labor, materials, capital, and other inputs. Because those markets are perfectly competitive, the supply price of each input to the firm is the given market price, regardless of how much the firm buys.
3. The firm uses only one variable input. Other factors are fixed during the analysis. For example, labor might be the variable input, while capital and land are fixed during the period being considered.

The firm will follow this profit-maximizing rule: Use the input up to the level at which the added *cost* from one more unit of the input just equals the added *revenue* from the output that the last unit of input produced. ***More precisely, the firm uses the input at the level where the cost and revenue of the marginal unit of input are equal.*** The key comparison is between cost and benefit at the margin.

To see the decision clearly, we need to discuss its parts one by one. Half of the choice rests on the cost of the input. Since the firm is a price taker in the input market, it can buy as much of the input as it wants at the going market price. Every time it uses another unit of input, therefore, the addition to cost is simply the price of the input. Input price is often referred to as the marginal cost of the input. However, do not confuse the marginal cost of the input with that of output. They are not necessarily the same. For example, if a unit of input costs \$10, that is the marginal cost of the input. If that unit of input can produce two more units of output, the marginal cost of the output will be \$10/2 or \$5.

Marginal revenue product

The other half of the comparison is ***marginal revenue product (MRP),*** the dollar value of the output produced by the marginal unit of the input:

$$\begin{pmatrix}\text{Marginal}\\ \text{revenue}\\ \text{product}\end{pmatrix} = \begin{pmatrix}\text{Marginal}\\ \text{product}\\ \text{of the input}\end{pmatrix} \times \begin{pmatrix}\text{Marginal revenue}\\ \text{of output}\end{pmatrix}.$$

The relationship between the quantity of an input and its marginal revenue product is called the *MRP curve.* The shape of the MRP curve depends on both the MP and MR curves. The marginal product schedule may at first slope upward, but will then

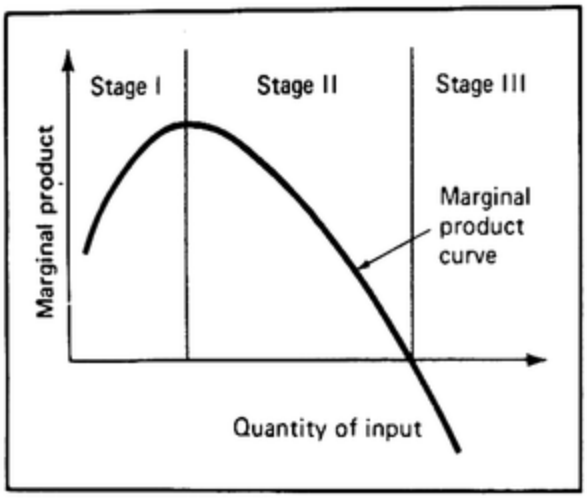


Figure 1 Three stages of the marginal product curve

Stage I is associated with rising marginal product, Stage II with declining marginal product, and Stage III with negative marginal product. Stage I is characterized by too little of the variable input relative to the fixed input for efficient production. The *fixed* input is being wasted. Stage III is characterized by too much of the variable input relative to the fixed input for efficient production. The *variable* input is being wasted. Only in Stage II are both inputs being used in efficient amounts relative to each other. Therefore, profit-maximizing firms will only produce in Stage II, where marginal product is declining.

reach a maximum and begin to decline, as in Figure 1. The eventual decline of the marginal product curve is explained by the *law of diminishing returns*. If a firm uses a fixed input and a variable input, at some point further additions of the variable input will add less and less to output. A firm will normally want to operate in the area of diminishing marginal productivity.

As Figure 1 shows, the marginal product schedule may include three stages. In Stage I, marginal product is rising. In Stage II, marginal product is declining, and in Stage III it is negative. There seems to be something good about increasing marginal productivity, so the intuitive reaction might be that firms will choose a level in the range where the marginal product is increasing. But in this case, intuition is wrong. With a little thought, you can see why.

Start by thinking about Stage III. Obviously, no firm will produce in Stage III, where the marginal product of the variable input is negative. There is so much of the variable input relative to the fixed input that more of the variable input actually makes total output decline. Now consider Stage I. While Stage III was characterized by too much of the variable input, Stage I has so little of it that much of the fixed input is wasted so that its marginal product may even be negative. While Stage III is associated with wasteful amounts of the variable input, Stage I is associated with wasteful amounts of the fixed input. Only in Stage II are both inputs being used in efficient amounts relative to each other. In the long run, therefore, a firm will only produce in Stage II, where the marginal product schedule is downward-sloping. This means that when the marginal revenue product schedule is being derived, only the downward-sloping portion of it is relevant in general.

The next question in determining the shape of a firm's marginal revenue product schedule is what the firm's marginal revenue schedule will look like. As you saw in earlier chapters, the marginal revenue schedule can take one of two general shapes. For a *perfectly competitive firm*, the marginal revenue schedule will be horizontal. For an *imperfectly competitive firm*, such as a monopoly or an oligopoly, the marginal revenue schedule will be downward sloping.

As Figure 2 shows, the marginal revenue product schedule for a competitive firm is the product of the downward-sloping portion of the marginal product schedule and a horizontal marginal revenue schedule. For a competitive firm, this is often called the ***value of marginal product*** schedule. For an imperfectly competitive firm, the marginal revenue product schedule is the product of the downward-sloping portion of the marginal product schedule

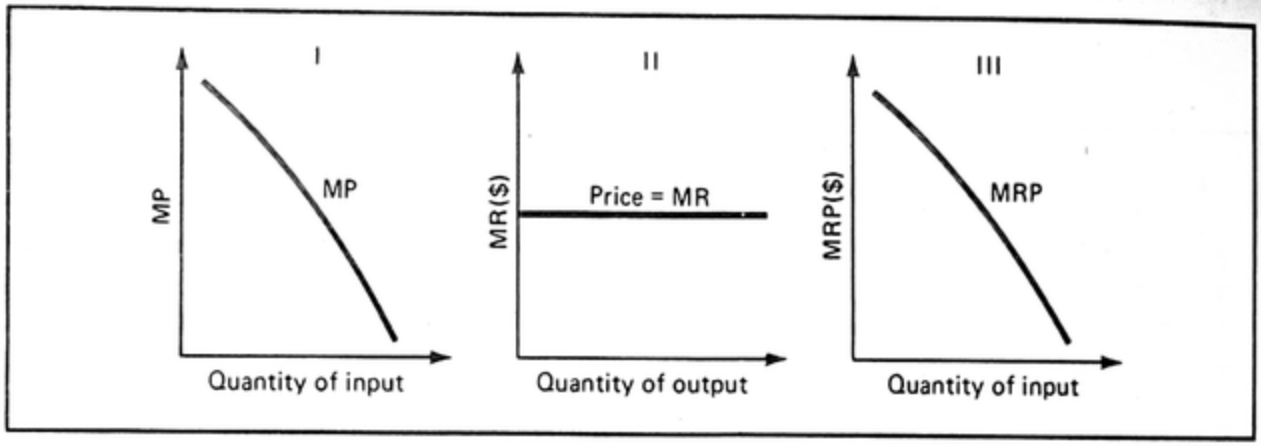


Figure 2 Deriving the marginal revenue product curve for a perfectly competitive firm

For a perfectly competitive firm, the marginal revenue product schedule is derived by multiplying the declining portion of the marginal product schedule by the constant marginal revenue. The result is a declining marginal revenue product schedule.

and a downward-sloping marginal revenue schedule, as shown in Figure 3. The result in both cases is a downward-sloping marginal revenue product schedule.

Now that both the change in cost and the change in revenue from using an additional unit of input have been derived, that information can be used to determine the profit-maximizing level of input use.

The profit-maximizing level of input use

In Figure 4, the price of the input has been added to the diagram of the marginal revenue product curve. To the left of Point *A*, each unit of input adds more value than it costs: MRP exceeds the price of the input. The blue shaded area shows the net addition to the firm's profits from hiring those inputs, whose value exceeds their costs. The firm will hire those units, since doing so increases its profits (or reduces its losses).

Beyond Point *A*, the input's MRP has fallen below the input's price. Since each added unit of input now costs more than it adds to the value of production, it causes net financial losses to the firm, as shown by the gray shaded area to the right of Point *A*. The firm will not use any inputs to the right of Point *A*.

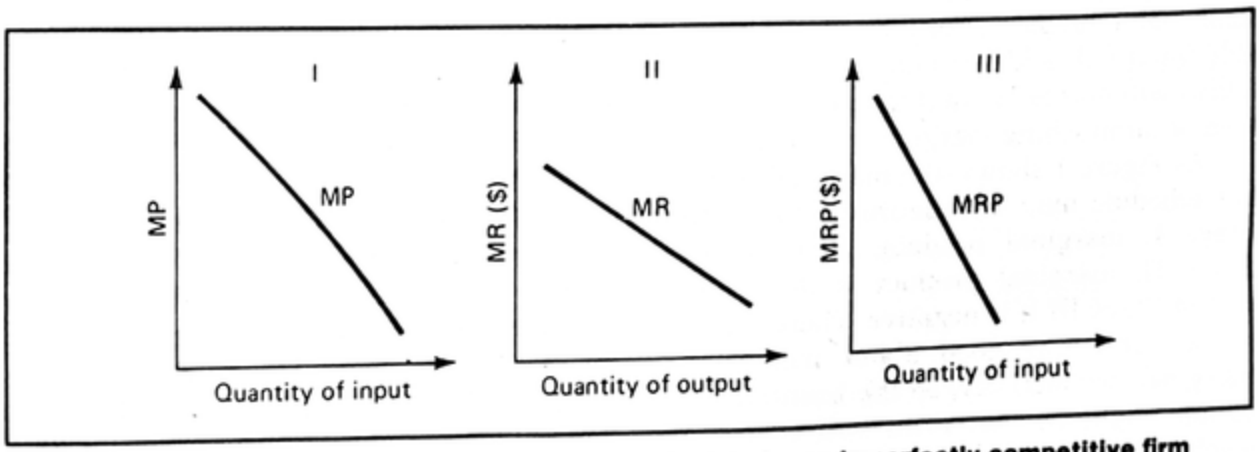


Figure 3 Deriving the marginal revenue product schedule for an imperfectly competitive firm

For an imperfectly competitive firm, the marginal revenue product schedule is derived by multiplying the declining portion of the marginal revenue product schedule by a declining marginal revenue schedule. The result is a declining marginal revenue product schedule. The marginal revenue product schedule will decline more sharply than it would if the firm were completely competitive.

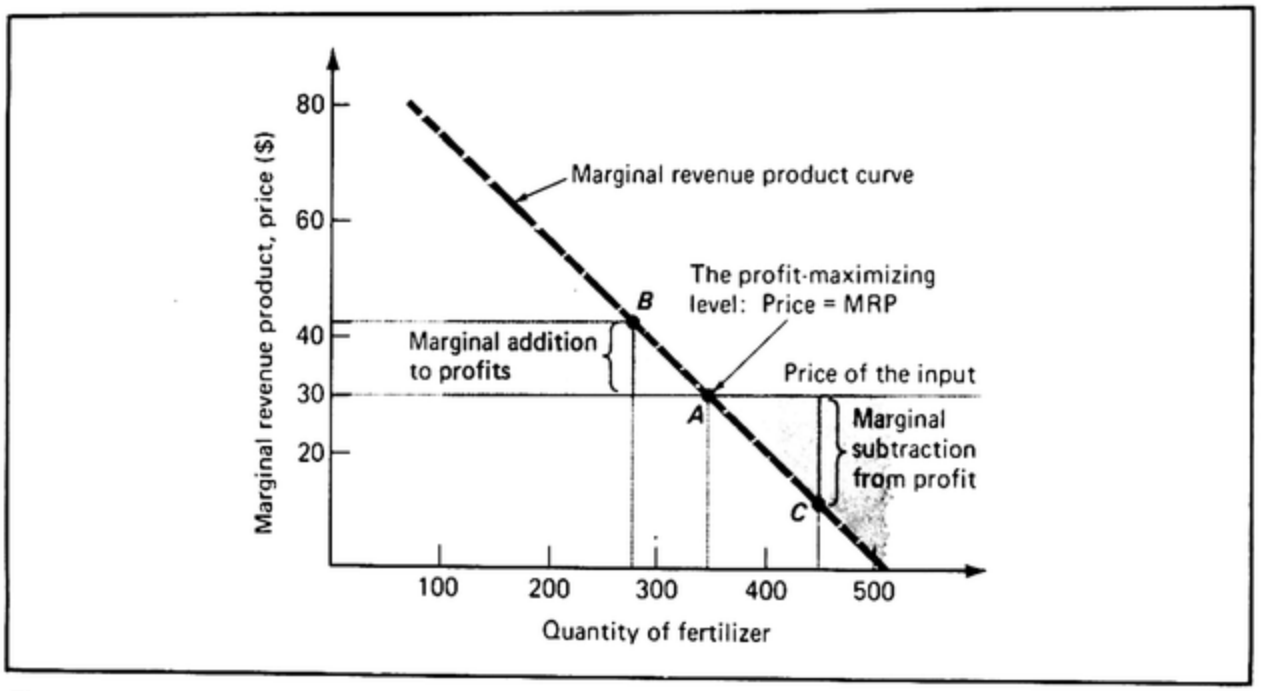


Figure 4 The profit-maximizing level of the input: *MRP* = input price

At Point *A*, the price paid for the marginal unit of input just equals its marginal revenue. At lesser amounts, such as *B*, the MRP exceeds the price. At higher amounts, such as *C*, the input's price exceeds its MRP. Profits will be reduced by using more inputs than at *A*.

At precisely Point *A*, the firm has gotten the maximum net profits that the input can provide, while avoiding the net losses: *The input's MRP equals its price.* Marginal value equals marginal cost for this input. That condition is precisely analogous to the marginal-revenue-equals-marginal-cost rule for the firm in maximizing its profits in choosing its *output* level.

Deriving the firm's demand schedule for an input

Using the profit-maximizing rule of marginal revenue product equals input price, the firm's demand schedule for the input can be derived. The analysis holds only if we assume that the firm is a price taker in the input market, and uses only one variable input.

A firm's demand schedule for an input will show the amount of the input that the firm will wish to purchase at different input prices. Taking the input price as given, the firm will wish to purchase the profit-maximizing amount of the input. In Figure 5, for example, at a price of $10 per unit of input, the firm would want to adjust the quantity of input it uses until the MRP equals the input price of $10. The firm would wish to purchase 170 units of the input at a price of $10, and Point *A* would represent one point on the firm's demand schedule for the input. To generate additional points on the firm's demand schedule for the input, simply vary the input price and locate the quantity that will equate marginal revenue product and price. At a price of $25 per unit of input, the profit-maximizing quantity of input would be 130 units. That price-quantity combination of $25 and 130 units of input, Point *B*, would represent an additional point on the firm's demand schedule for the input. At an input price of $45 per unit, the profit-maximizing quantity of input would be 80 units. Point *C* would be yet another point on the firm's demand schedule for the input. As the input price changes, the new price-quantity combination will always lie at some point on the firm's marginal revenue product schedule. The firm's demand curve for the input is

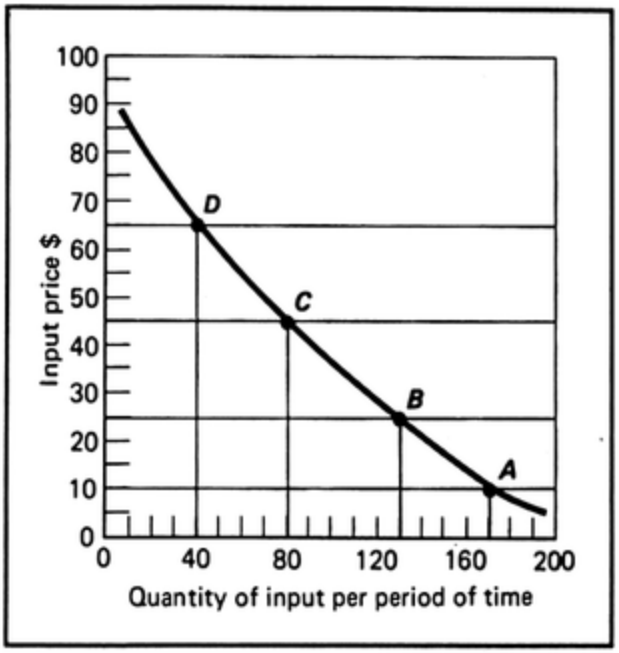


Figure 5 Derivation of a firm's demand schedule for an input

simply the marginal revenue product schedule.

The demand schedule for inputs also slopes downward. This implies that a firm will find it profitable to use more of the input as its price falls. In Figure 6, as the price of a variable input falls, the firm's marginal cost schedule shifts to the right, reflecting the lower costs of production. The profit-maximizing point, the new marginal revenue–marginal cost intersection, will shift from Point *A* to Point *B*. The firm will now find it profitable to produce more output.

If more output is to be produced, more input must be purchased. Thus, the demand schedule has a downward slope, indicating that the quantity of inputs demanded will rise as their prices fall.

Elasticity of demand for an input

The extent to which the quantity demanded of the input will respond to a change in input prices can be measured by the elasticity of demand for the input. Three conditions make the demand for the input relatively elastic or responsive to changes in input prices:

The input accounts for a large percentage of total costs Suppose that the price of an input falls by 50 percent. That sounds like a big change. However, the impact on quantity demanded when the input accounts for 2 percent of total costs will be quite different from when it accounts for 70 percent of total costs. The higher the percent of total costs the input accounts for, the more important it is in the total cost picture and the bigger the shift in marginal cost from a given percentage change in input price. This will make the profit-maximizing level of output and, therefore, the demand for input more responsive to changes in input price.

The demand for output is relatively elastic As input price falls and the marginal cost schedule shifts to the right, the profit-maximizing price of output will also fall. The more elastic the demand for output, the larger will be the increase in the profit-maximizing quantity that will accompany a given change in price. The larger increase in the demand for output will cause a larger increase in the demand for input for that given price change. Therefore, the more elastic the demand for output, the more elastic the demand for an input will be.

Substitution is easy A third condition is the ease of technical substitution. In the long run, a fall in the price of the variable input will cause a substitution of the variable for the fixed factor. If the technology of the firm allows one factor to be easily substituted for another, there will be relatively larger change in the demand for the input when its price changes.

However, substitution of variable for fixed factors cannot explain elasticity along a given input demand schedule. The

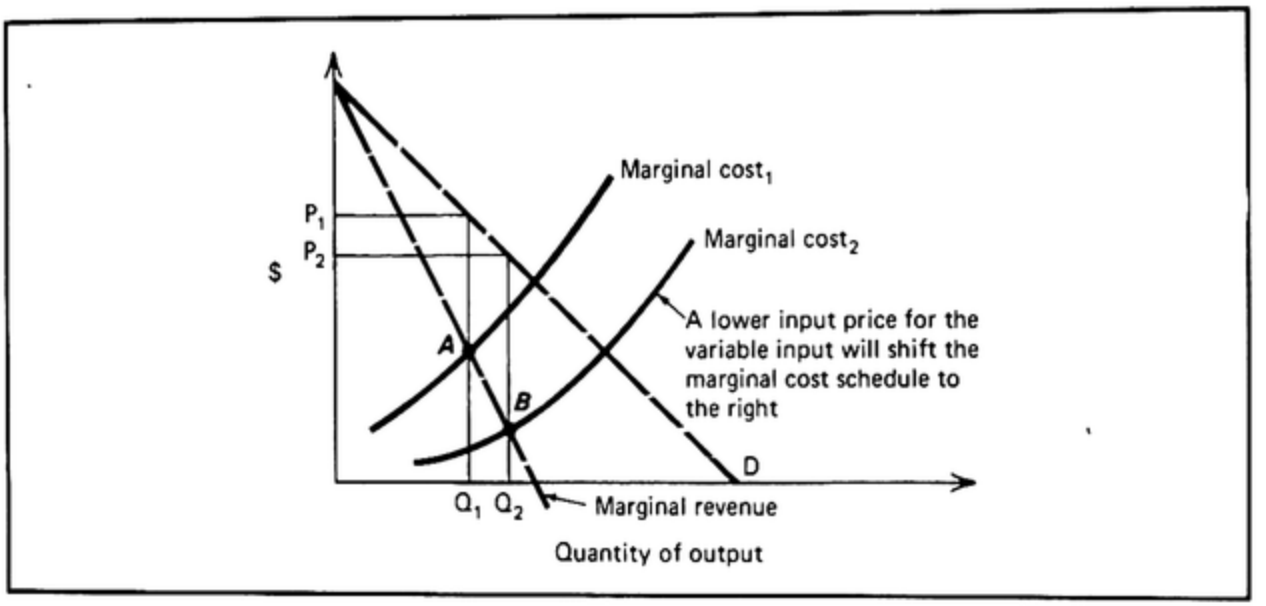


Figure 6 A change in input price will change the profit-maximizing level of input use

Originally, the firm finds it profit maximizing to produce Q_1 at a price of P_1. When the price of the variable input falls, the marginal cost schedule shifts right. The profit-maximizing point is now *B*, which represents a higher level of output, Q_2. To produce the higher level of output, more of the variable input must be bought. This explains why the demand schedule for the input is downward sloping, with more input being purchased at lower input prices.

reason is that as such substitution takes place, the marginal product schedule, from which the demand schedule for the variable input was derived, will also shift, as a result of changes in the quantities of other inputs. Ease of technical substitution only helps to determine the overall change in the demand for the input as its price changes, not elasticity along a given demand schedule with other inputs fixed.

Shifts in the marginal revenue product schedule

Since the marginal revenue product schedule is derived from the marginal product and marginal revenue schedules, it will shift only if either the marginal product schedule or the marginal revenue schedule also shifts. For example, if the demand for output changes because of a change in consumer income, taste, or population, the marginal revenue schedule for output and, therefore, the marginal revenue product schedule must shift as well. If the marginal product schedule shifts because of a change in technology, the marginal revenue product schedule will also shift. Remember that a change in the price of the variable input will simply cause a movement along the existing input demand schedule, at least in the short run. In the long run, a change in the price of the variable input may cause the input demand schedule to shift because of changes in the amount of the fixed factor used, which will cause a shift in the marginal product schedule.

Comparison of input use of a perfect competitor and a firm with monopoly power

As you have seen, the marginal revenue product schedule is derived from both the marginal product schedule and the marginal revenue schedule. For a perfectly competitive firm, the marginal revenue schedule is a horizontal line at the level of market price. For a monopolist, the marginal revenue schedule slopes downward and lies below or to the left of the firm's

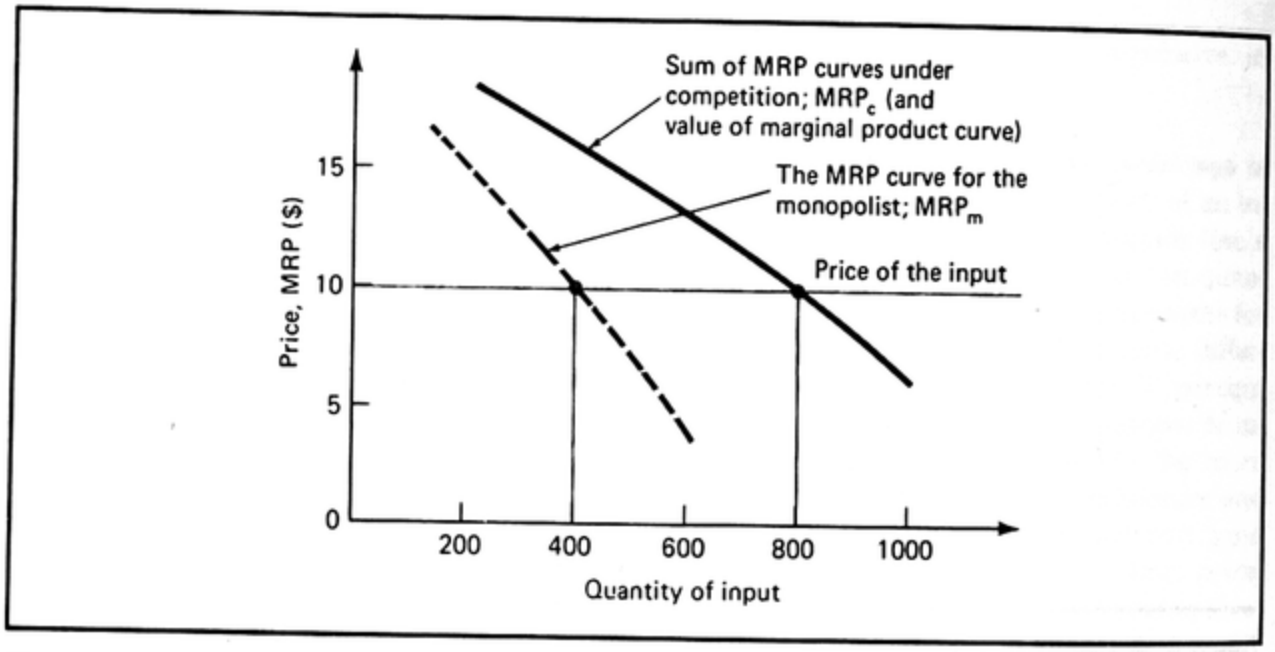


Figure 7 The monopolist's MRP curve lies below the summed competitive MRP curve

The sum of the competitive firms' MRP curves is shown by *MRP_c* (c for competitive). This is also called the value of marginal product curve. But the monopolist has a marginal revenue curve lying below its demand curve. Those MR values are used (rather than competitive price) to derive the *MRP_m* curve, which therefore lies below the *MRP_c* curve.

demand curve. These differences in the marginal revenue schedules of perfect competition and monopoly lead to important differences in the allocation of resources.

In Figure 7, MRP_c represents the sum of the marginal revenue product schedules of the firms in a perfectly competitive industry. MRP_m represents the marginal revenue product schedule for an industry in which the firm or firms have monopoly power. Since the marginal revenue curve for such an industry would lie below the demand curve, its marginal revenue product curve will lie below the competitive marginal revenue product schedule, which would coincide with the industry demand curve. The resulting competitive schedule is often referred to as value of marginal product. Since the marginal revenue product schedule represents the firms' demand for inputs, you can see that monopoly power reduces the amount of inputs that firms in the industry will wish to buy at a given input price. In Figure 7, competitive firms will purchase 800 units of the input at a price of $10 per unit. The monopolistic firm will purchase only 400 units of the input at the same price. The reason is, of course, that the downward-sloping marginal revenue schedule of the monopolist, which lies *below* the demand schedule, reduces the value or return to the firm of an additional unit of input.

The key result is that *monopoly power reduces the amount of inputs used by an industry*, compared to the amount of inputs used under competition. This reduction in input use is consistent with the monopoly restriction of output that was explained in Chapter 10. Obviously, if less output is to be produced under monopoly, smaller amounts of the input are needed. Note, too, that the reduction in production of output and in input use under monopoly is not due to any conscious decision on the part of the monopolist to restrict output. Both a competitive firm and a monopolist

compare marginal costs and marginal benefits in determining production levels. Both set marginal revenue equal to marginal cost to determine output levels, and marginal revenue product equal to the price of the input to determine input use. Because of the differences in the marginal revenue schedules, however, the same rules lead to quite different results.

Taken altogether, the analysis also provides the link between consumers' final demand and the firms' demand for inputs. Consumers express their preferences and spending power in their demand curves, which, working with supply, set the market price. Those market values then transmit back to give MRP its specific values. If consumers' demand rises, that will usually increase the output's price and cause MRP to shift up. Final demand, therefore, influences the demand for inputs. For example, when consumers' preferences change to smaller cars or more formal clothes, then the demands for inputs to make those goods will rise. Such adjustments are routine, linking all outputs and their inputs.

Recall that the demand for inputs is a "derived demand," arising from the final demand for goods. We have now shown how the input demand is derived. It may proceed back through many stages—for example, from a refrigerator to the sheet steel for its surface, to iron, to the iron ore and coal used to smelt it. At each point, the same link exists between the demand for the firm's output and the firm's demand for its inputs.

Supply and equilibrium in input markets

Now we turn to the supply of inputs and to the equilibrium results that occur in input markets. First we explain why input supply curves slope up. Then we discuss the supply-demand equilibrium and, finally, economic rent, a key concept.

The supply of inputs

At the market-wide level, inputs conform to the general rule that *supply curves slope up*. But the causes differ from those in output markets, where supply curves reflect up-sloping marginal cost curves of firms.

To understand input supply, remember that the individual competitive firm faces a horizontal supply curve of each input. The firm buys such a tiny share of the whole market's supply that its own actions do not affect the input's market price.

But for the whole market, the supply curve is rarely horizontal. Instead, increasing amounts usually can only be obtained at higher prices. There are two main reasons for this up-slope in the supply curve.

1. Opportunity Cost: The input must be bid away from valuable alternative uses To get larger amounts, higher prices must be paid to draw the input from increasingly valuable alternatives.

For example, more trained mechanics may be needed. The first 1,000 of them can be obtained by offering salaries of $20,000 per year. To obtain the next 1,000 mechanics, however, it may be necessary to offer annual salaries of $25,000 to get them to leave well-paying jobs in other industries. The next 1,000 must be attracted away from specialized aerospace jobs that pay $30,000. Therefore, they must be paid at least $30,000. This feature alone would cause the supply curve to slope up.

2. Direct Cost: It may be increasingly costly to produce the input Small amounts may be obtained by simple, cheap processes, but larger amounts may require expensive methods of production. Therefore, the supply curve may slope up to reflect these increasing direct costs of production.

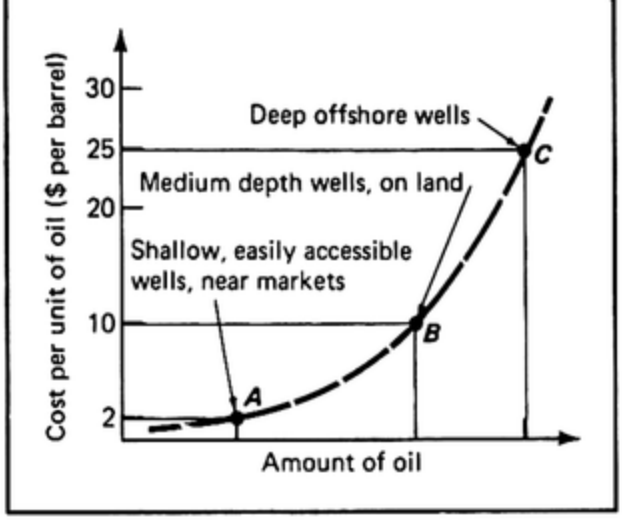


Figure 8 Increasing direct costs can make an input's supply curve slope up

The cheap, easy sources are tapped first, as at Point *A*, where oil only costs $2 per barrel to find and extract. To get the amount at Point *B*, wells must be drilled deeper, in less accessible locations. Costs are therefore higher—$10 per barrel. Obtaining the amount of oil at Point *C* involves much higher costs, for expensive rigs to drill deep wells out in the ocean. A price of $25 or higher will induce companies to incur those costs.

The best examples of this are natural resources, such as coal, ores, water, and fertile land. Each exists in a variety of qualities and locations. Some are easy and cheap to use: easily accessible oil, water flowing in nearby rivers, thick coal seams near the surface, the best-quality land nearest the cities. They are used first, at low cost, as shown for oil by Point *A* in Figure 8. Then, to get more supply, the more difficult sources must be tapped, at the higher costs shown by Point *B*. Still higher costs are incurred for still higher quantities, at Point *C*, which represents very deep offshore wells. Since the costs of extracting the oil are so high at this point, oil will be supplied only at high prices.

The point also applies to other resources—for example, to the recruitment of a specialized labor force to operate a new enterprise. Training the necessary labor may be increasingly costly. The first trainees may well have innate talent, and so their training will be rapid and cheap. Further candidates are likely to be less talented and therefore costlier to train.

These two conditions—opportunity costs and rising production costs—give the supply curves their characteristic positive slope. The degree of slope will vary from case to case, but the same logic applies to them all. If the firm operates in both a competitive output and an input market, the contribution of each input to the process of production is easy to determine. In that case, each input is paid an amount equal to the value of its marginal product. As you know from the earlier discussion in this chapter, the value of marginal product represents the inputs' marginal contribution to revenue. The input payment is set at precisely what the input is worth to producers. Therefore, competitive markets tend to pay inputs what they contribute to revenue at the margin.

Market equilibrium As in any market, the equilibrium price and quantity for each input are determined by the interaction of the supply and demand forces. Figure 9 illustrates the equilibrium result for one input market. The demand and supply

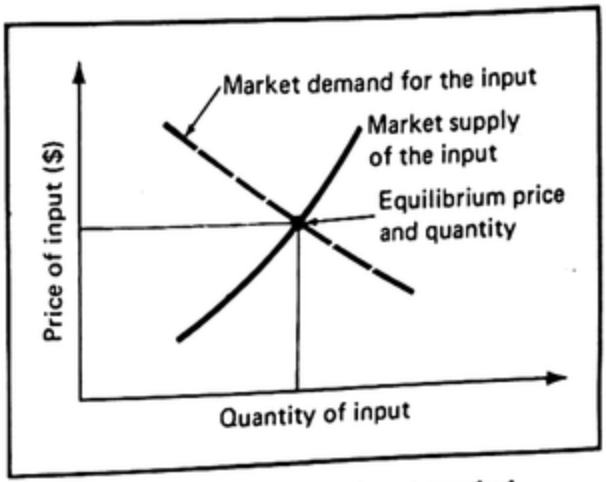


Figure 9 Equilibrium in an input market

As with output markets, the equilibrium price and quantity in input markets are determined by the interaction of supply and demand forces. While the demand and supply schedules for an input have the same slopes as do the demand and supply schedules for output markets, remember that the explanations for the slopes differ markedly.

schedules for the inputs have exactly the same basic appearance as do the demand and supply schedules for output markets. But remember that the explanation for the slopes of the schedules, and for the derivation of the curve, is different for input and output markets.

Economic rent

One special feature of supply is ***economic rent.*** To explain it, we begin with the polar case of perfectly inelastic supply. Then we show how economic rent occurs when supply has varying degrees of elasticity.

Perfectly inelastic supply When a good's supply is perfectly inelastic—with a vertical supply curve—the same quantity of it will be supplied regardless of the price. The price depends strictly on the level of demand, as in Figure 10. The amount, Q_1, will be supplied when the price is as low as P_1 or even zero. If demand shifts up, the price will rise (as to P_2 and P_3), but quantity will stay at Q_1.

In such a case, all payments to the input's owners are *economic rent.* They are not cost or profit. ***Economic rent is a payment in excess of the price needed to elicit supply.***

Rent is common for inputs, especially natural resources. Urban land is the economists' traditional instance of perfectly inelastic supply and, therefore, of pure economic rent. Each plot of land in a city is merely an area upon which buildings and valuable activities can be located. The economic uses of such land generate value, which, in turn, give rise to demand for the land itself. As the density of economic activity rises, the demand for the land also rises. That causes the price of the land to rise, as to P_2 and P_3 in Figure 10.

Therefore, urban land prices (which provide economic rent) reflect economic density. Within a city, the more densely

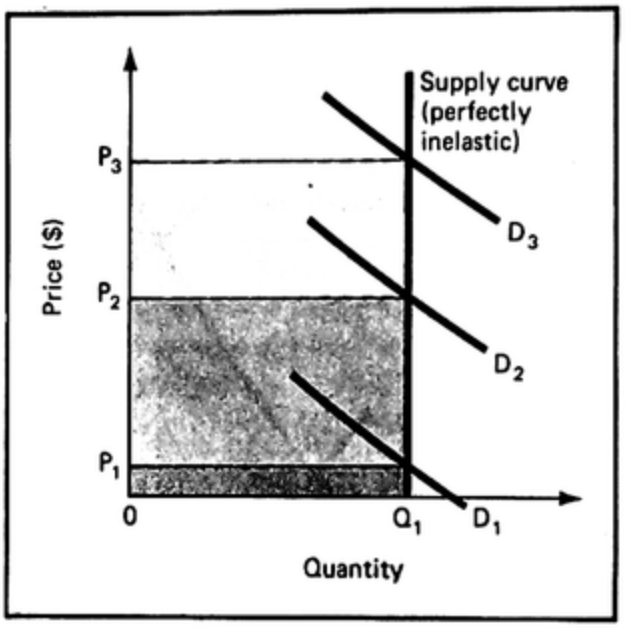


Figure 10 When supply is perfectly inelastic, all payments are economic rent

Because the same quantity, Q_1, will be supplied at all prices, such as zero, P_1, P_2, or $P3$, the actual price is determined by demand. All payments are economic rent; the total rent payments for each price (P_1, P_2, P_3) are shown by the shaded areas below those prices.

used central land costs more than peripheral land. Over time, a rising density of use will raise prices (and economic rent) on all land.* Each city has contours of land values, with high prices at the center and lower prices at the edges.

The analysis applies to any good that is in fixed supply. All payments to it are economic rent.

When supply has a degree of elasticity In general, supply is up-sloping but not perfectly inelastic. The payments to the input provide both *economic rent* and *transfer earnings.* ***Transfer earnings are payments that are necessary to elicit supply;*** in competitive markets, *transfer earnings are identical to cost.* For example, with lumber, in Figure 11, the supply curve slopes upward,

*Economic rent is not to be confused with "rent" for leasing apartments, automobiles, and the like. A rental rate is merely a periodic payment for a service. It need not contain any element of true economic rent.

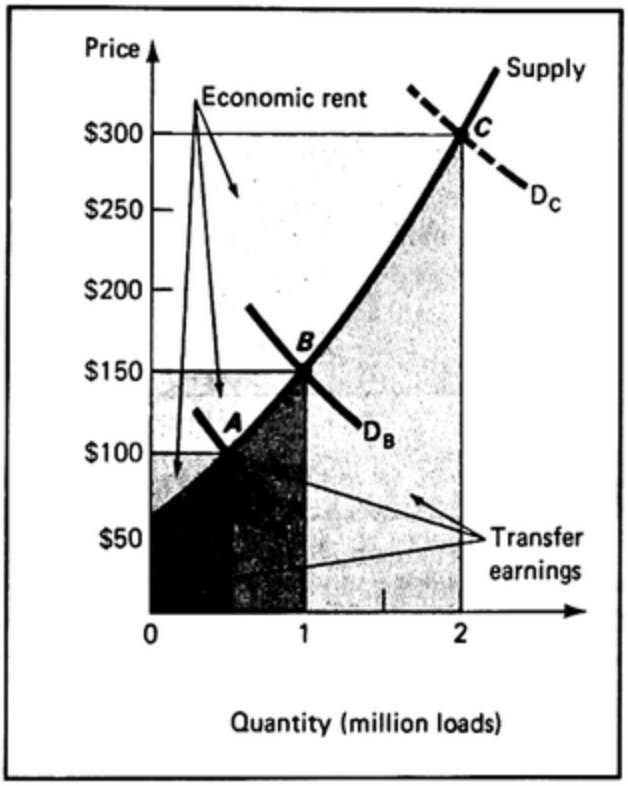


Figure 11 Separating economic rent and transfer earnings

The height of each point on the supply curve represents the dollar payment required if the lumber is to be supplied. Summing all such heights, the area under the supply curve represents total transfer earnings. The areas *above* the curve but *below* the going price represent economic rent. Transfer earnings are shown below the curve, for various prices; economic rent is shown above the curve for various prices.

starting with the most accessible and suitable areas for harvesting lumber (at Point *A*), and then moving to less and less accessible and productive forest land (at Points *B* and *C*). Successively higher costs of harvesting the trees must be incurred as the quantity of lumber increases. Therefore, higher prices must be offered to lumber suppliers to elicit more lumber.

Any payments above those transfer earnings are economic rent. For example, if a particular load of lumber will be supplied for $100, but the going market price is $300, then the supplier of that load receives $200 in economic rent.

Figure 11 shows how the total payments to this input can be divided between transfer earnings and economic rent. Suppose that the equilibrium market price-and-quantity combination is $150 per load and 1 million loads, as at Point *B*. The supply schedule shows that the 500,000th load will be offered at a price of $100, which is its cost of production. Because market price is $150, the 500,000th load receives transfer earnings of $100 plus economic rent of $50. The 750,000th unit will be supplied only at a price of $125, and so the market price of $150 provides it with $125 in transfer earnings and $25 of economic rent. The 1 millionth unit (supplied at the going price of $150) earns no economic rent at all; the $150 is all transfer earnings.

In general, since the height of the supply curve shows the cost of supply, the area below the supply curve represents transfer earnings, which cover the suppliers' costs. The area that is above the supply curve and below the horizontal line representing the market price is the total economic rent received at that market price. It is shaded blue in Figure 11. Some units obtain large economic rent, while others get little, and the last unit supplied receives no rent at all.

If demand increases to level D_C the equilibrium price will rise to $300, as shown. The quantity supplied will rise to 2 million. As before, the last unit supplied (now the 2 millionth) receives only its transfer earnings, with no economic rent. But the price rise enlarges the economic rent gained by other units. The 500,000th unit's economic rent rises to $200, the 750,000th unit's rent to $175, and so on. The total additions to transfer earnings and economic rent are shown by the shaded areas in Figure 11.

When demand rose to D_C, the original suppliers of the first 1 million loads did not change their choices or behavior. But

their economic rent rose substantially. Such rises in economic rent are often called ***windfall gains,*** for they occur with no added effort or contribution by the owners of the inputs. Recent examples have come from the rises in real estate and in oil prices during the 1970s. The holders of those assets simply gained extra value without altering their own decisions or activities.

Taxing economic rents

Land is both a fixed factor, which earns economic rents, and a large share of the value of all assets. Therefore, it has always attracted interest as a source of taxes. In the pure case, since its supply curve is vertical, all payments to it are economic rent, as in Figure 12.

If a tax were set to take Area I, or even the entire shaded area, the land would still be supplied and used. The price paid by the users would be as high as before (P_1), reflecting the land's economic value as an input. But the state would simply take some or most of the owners' rent. Nothing in the economy's efficient production choices would be changed, and the government would have more tax revenues. Moreover, the tax is easy to apply, for the land is immobile and visible.

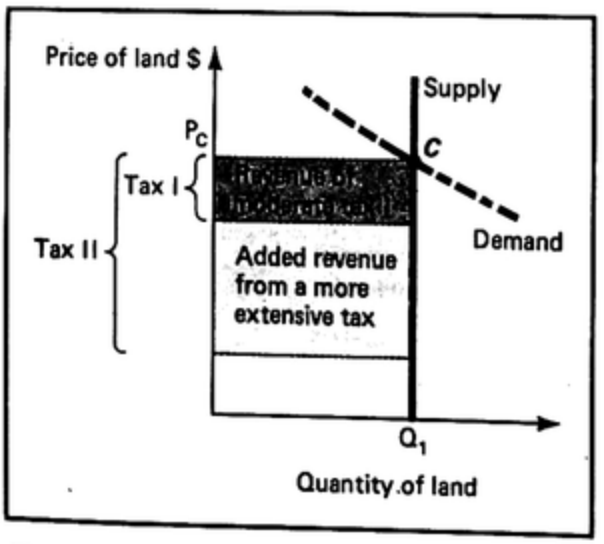


Figure 12 Taxation of rent from land

If land is a fixed factor, its supply curve is vertical, as shown. This would be likely for land in towns, where the amount in a given area is fixed. The equilibrium price is P_c. If Tax I is levied, the government will receive the revenue shown by Area I. The steeper Tax II would yield the total shaded area. Yet, even though the owner would be receiving a much lower net payment for the land, he would still supply the land in the same amount, Q_1.

These points became the heart of a crusade in the 1880s by an amateur economic thinker named Henry George. A tax on land would make an ideal source of revenue, he urged. Land values rose merely because cities grew. Since landowners contributed nothing to deserve that rise in rent, he argued that taxing away this newly created rent was both efficient and fair. The tax might even cover all of government's costs, at a time when public spending was far less important than it is now. Hence it was called a "single tax."

An effective orator, George was nearly elected mayor of New York in 1886. But a conservative reaction set in and George's single-tax movement faded. Nonetheless, a Henry George Society still exists in New York, and economists have always accepted the core logic of his argument. Indeed, his argument was derived directly from the marginal utility theory, which economists were developing at that time.

However, the single tax does have practical defects. As we noted earlier, there is usually a mingling of land and building values. Since both rent and cost are often present, separating them is usually difficult, even though in principle it could be done. Another practical limit is that nowadays the land tax would not be able to provide all of the national budget. Land's relative importance has dwindled, as capital has expanded and government budgets have multiplied a hundredfold. The "single tax" could no longer cover much of total public spending.

Yet, the tax on economic rent *has* become a major part of modern public finance, as local property taxes. Levied by virtually every city, these taxes on residences and businesses are usually set at between 0.5 and 1.5 percent of the value of property. The revenues are usually over half of the city budgets.

Who provides the value of production?

Now we advance to one of the most divisive economic issues in the history of economics: Which factor provides the most value in the production process?

The basic problem is readily apparent. When a person spins wool by hand and then knits a sweater from it worth $50, the labor has provided most of that $50 value. But complex modern production is not so easily dissected. Consider a row of gleaming new $8,000 cars emerging from an assembly plant. Labor, capital, and land were used together to produce the cars, and each provided some of the $8,000 value. But exactly *how much* of the value did each factor add?

That question poses an explosive issue, for each factor's owners naturally think that theirs is the most important one. Economists have debated the issue for over three centuries, and their diverse views are shown in Table 1. Gold, land, industry, capital, and labor have each been credited at some time with being the main source of value.

The issue has great importance in labor negotiations. Workers regard their labor as the main source of value and claim high wages as their due. The company managers see their factories and equipment as the crucial factor of production and resist wage increases to gain higher profits on the invested capital.

The issue arises even in your classrooms. Whose efforts cause learning to occur? Teachers often feel that their lectures, labs, and office hours have instilled the knowledge. Students instead often give the credit to their own hard work.

In such cases, all inputs contribute, but how much? If all markets are competitive, then each firm buys its inputs at competitive prices. Therefore, it is paying

Table 1 Alternative views about the most productive factor

Person or Group	Main Period	Primary Source of Economic Productivity
Mercantilists (Western Europe)	1650–1750	Gold
Physiocrats (France)	1750–1780	Land (in agriculture)
Adam Smith (Britain)	1770–1789	Industry and trade; also land to an extent
Industrial spokesmen "Manchester Liberal" economists (Britain)	1800–1890	Capital
Austrian capital theorists		
Karl Marx (Western Europe)	1850–1883	Labor (capital merely embodies the labor that made it)
Neoclassical economists (Western Europe and U.S.A.)	1870–present	All factors share in productivity according to their marginal revenue product

each factor an amount equal to its marginal revenue product. Since MRP is each factor's contribution to production, *the payment for each factor is set precisely at the economic value that it adds to production.*

Under competition, these payments to the factors use up all of the firms' sales revenue, so that there is no surplus money that could be given to any factor. A well-functioning competitive market system, therefore, tends to pay all inputs approximately what they contribute to production at the margin. Input choices are efficient, and the division of payments among inputs has a definite basis.

Yet, this basis is narrow, and what is efficient may not be fair. The narrowness and possible unfairness will be discussed later, when we present the general equilibrium outcomes for the whole economy. The effect of monopoly on labor incomes will be considered sooner, in the next chapter.

Summary

This chapter deals with the demand for and supply of inputs. Its main points are the following:

1. The theory of input markets is based on three assumptions: the firm is a profit maximizer; the firm is a price taker in the input market; and the firm uses only one variable input in addition to its fixed inputs.
2. The addition to cost from using an additional unit of input is the price of the input.
3. The addition to revenue from using an additional unit of input is the product of the addition to output resulting from the additional unit of input (the marginal product) and the addition to revenue from the sale of this output (the marginal revenue). The product is called marginal revenue product.
4. A firm will use an input up to the point at which the last unit of input adds as much to cost as to revenue. This is the level of input use at which the price of the input equals marginal revenue product.
5. The firm's demand schedule for an input is the marginal revenue product schedule for the input.
6. The demand for an input will be more elastic: The larger the percentage of total costs that the input accounts for, and the more elastic is the demand for output.
7. A monopolized industry will purchase less of an input than would be the case if the industry were competitive.
8. While an individual firm is assumed to be a price taker in the input market, increased supplies of the input to the *industry* are likely to be available only at higher prices.
9. *Transfer earnings* are the minimum payment that the owner of the input must receive if the input is to be offered for sale. *Economic rent* is the payment to the owner of the input over and above what is necessary to prevent the owner from transferring that input to another use. An increase in the economic rent paid to the owner of the input is called *windfall gains.*
10. If a firm operates in output and input markets that are both competitive, each input's contribution to output is measured by its value of marginal product.

11. A tax on economic rent does not affect the supply of an input offered for sale.

Key concepts

Marginal revenue product (MRP)
Value of marginal product
Economic rent
Transfer earnings
Windfall gains

Questions for review

1. Which of the following statements is true? Explain your answer carefully.
   a. For one variable input, the price of the input and the marginal cost of output are the same.
   b. If marginal revenue product is greater than the price of the input, the firm must be making a profit.
   c. The profit-maximizing level of output must occur at the point where the marginal revenue product equals input price.
2. Indicate which of the following changes will cause a firm's demand schedule for an input to shift.
   a. The firm purchases more of the input.
   b. The price of the input falls.
   c. The price of a substitute for the firm's output falls.
3. Explain why monopoly power will influence the allocation of resources to a particular industry.
4. Why is a tax on pure economic rent more desirable than a tax affecting transfer earnings, from the point of view of resource allocation?

15

The Economics of Labor and Unions

As you read and study this chapter, you will learn:

- how people choose their jobs and their amounts of work
- how wage and employment levels are determined in the market
- the varieties of occupations and pay rates
- how monopoly elements (unions and monopsonies) alter the competitive wage and hiring levels

According to the book of Genesis, human life began in idyllic surroundings. In the Garden of Eden, the soil was rich, food was abundant, and the living was easy. After yielding to temptation, however, Adam and Eve were banished from that land of plenty, doomed to wrest their food from the soil with suffering, and to eat their bread with sweat on their brows. Simply put, they now had to work for a living. Unfortunately for us, this curse has extended to all of their offspring. Since that day, work has been the lot of humankind.

Economists have always recognized labor's great importance in the economic process. Work absorbs a large share of most people's efforts, time, and emotions. Work is not only how people "make a living." It also defines much of each person's success and sense of personal worth.

Labor is also a prime productive force in the economy, applied in millions of factories and stores. Like any other commodity, labor is bought and sold every day. There are many types

and grades of labor, all being sold at market prices. The whole economic process allocates labor, as people choose jobs, employers hire workers, and wage rates adjust.

Yet, because labor directly affects human welfare, it is not just another commodity like gravel or zinc. If you keep labor's special importance in mind, you will better understand the urgency of its social role. Policies toward labor—especially toward minimum wages, unions, and job discrimination—evoke strong reactions, and rightly so. They matter because work matters.

In the first main section of this chapter, we analyze the supply of and demand for labor. This is based on marginal productivity theory. The second main section discusses the varieties of labor. Finally, the third section analyzes departures from competitive conditions: unions and their effects, and employers' market power.

Labor supply, demand, and market outcomes

The basic unit of labor is the hour of work, in which skills and/or force are applied as part of a production process. The degree of effort and skill is assumed to be constant, so that each kind of labor-hour is a standardized input. Labor-hours are bought and sold on labor markets, with outcomes that are determined by supply and demand.

On the supply side, people have to decide whether to work at all, at what job, and for how many hours. On the demand side, firms have to decide how many labor-hours to buy. As these two sides of labor markets interact, the wage rates and quantities of labor hired are determined throughout the economy.

The marginal utility of work

Work is productive effort, done for a reward. Some work is interesting and satisfying, a source of *utility*. But work can also be hard, unpleasant, and, often, boring. The negative side of work is called *disutility*—the opposite of utility, or pleasure. Disutility embraces all the things that can make a job unpleasant.

The task is to explain (1) how people choose among jobs, and (2) how much work they will choose to do in those jobs at varying wage rates. The decisions depend largely on how much satisfaction or dissatisfaction the work brings, and that, in turn, depends on two elements: the person and the job.

Panel I of Figure 1 shows how the nature of the job may cause different amounts of satisfaction. The three curves illustrate how one person may view three different jobs, apart from whatever wages are paid. The average hours worked per day are plotted on the horizontal axis; ***marginal utility*** (the addition to job satisfaction) from each additional hour worked is on the vertical axis. Curve 1 represents a truly unpleasant job, which causes displeasure (negative utility) from the first moment.

Curve 2 represents a job that would give some satisfaction or enjoyment to the person who does it, for at least the first five hours or so. But the law of diminishing marginal utility applies, and the additional hours worked bring disutility. Curve 3 represents a highly enjoyable job. Only after 15 hours a day does disutility set in. For each person, then, different jobs will yield different amounts of satisfaction.

The three curves can also be viewed from a different perspective. As in Panel II of Figure 1, the curves could represent the same job as viewed by three different people. Person 1 dislikes the job entirely; the second person gets some utility from the

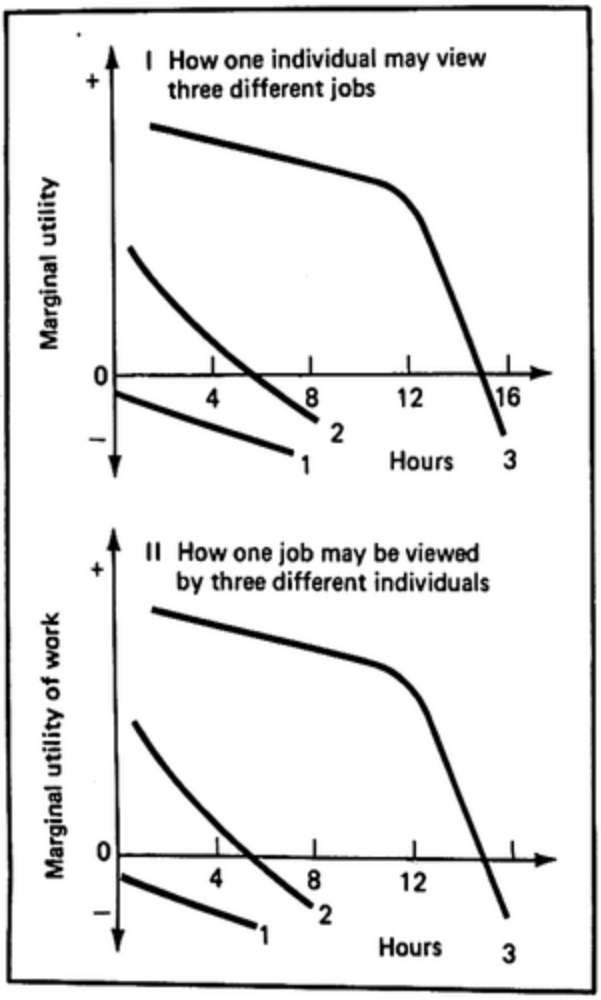


Figure 1 The marginal utility (and disutility) of work

Panel I represents the varying amounts of marginal utility (or satisfaction) that a person may derive from three different jobs. Job 1 brings no positive satisfaction for any of the hours worked. Job 2 satisfies up to the 5½-hour mark. Job 3 satisfies up to the 15th hour worked. Note that all of the jobs involve declining marginal satisfaction, with each hour worked bringing in smaller additions to satisfaction or larger additions to dissatisfaction.

Panel II shows that while different jobs may yield a person varying amounts of marginal utility, the same job may yield varying amounts of marginal utility to different persons. Person 1 receives only dissatisfaction from the job. Person 2 receives some positive enjoyment from the job, at least up to the 5½-hour mark. Person 3 enjoys the job the most, receiving increases in enjoyment for the first 15 hours worked. Note, however, that for each person, each additional hour worked causes smaller increases in satisfaction or larger increases in dissatisfaction.

job; while the third person would happily do it for 15 hours a day.

In all cases, even given the differing natures of jobs and people, the law of diminishing utility applies: There is decreasing satisfaction from additional hours of work. At some point, because people have to sleep, eat, and relax as well as work, extra work brings disutility.

Each rational worker aims to maximize the utility or satisfaction gained from a job by balancing the benefits of work (the enjoyment and the pay) against the cost (the disutility). Two decisions must be made simultaneously: which job to take and how long to work at it.

The choice of a job

Each person rules out many jobs because their disutility more than offsets the pay. People's job choices also reflect their own specific skills and preferences. Such talents can be innate or they can be acquired or developed through training. These skills are first discovered and developed in high school. Some students are attracted to specific jobs, such as auto mechanics, cooking, or carpentry, and may pursue vocational training or go directly to work. Those who attend college go through a further sorting process, deciding about future work on the basis of their interests and aptitudes. The job selection process leads people into the jobs that they are relatively best at—the jobs for which they have a *comparative advantage.*

Of course, in actually choosing a job, you have to consider not only your skills and interest but also how much it pays. You might get your greatest satisfaction from painting, but if no one will pay for your masterpieces, you will have to paint for a hobby and find another line of work. The wage rate you will accept depends largely on how interested you are in the job. The greater your interest, the lower the pay you may be willing to accept.

But wage rates also affect the *number of hours* people work.

Individual labor supply schedules

Each person has a supply curve of labor showing the amount of work she or he would choose to do at differing rates of pay. Normally this curve will slope up, as it does in Figure 2, because a higher rate of pay is necessary to overcome the increasing marginal disutility of work.

As Figure 2 shows, the elasticity of labor supply may differ from one person to another. Bentley has a more elastic supply curve than Angelos. A wage increase from \$7 to \$10 makes Bentley willing to work 3½ hours more, while Angelos would only be willing to work for another hour. A person whose time available for work is limited, such as a parent of young children or a full-time student, would tend to have a relatively inelastic supply schedule for labor. An increasing wage rate would not induce that person to work many more hours.

Some jobs have fairly inflexible hours: You must work eight hours a day or refuse the job. Other jobs offer more flexibility.

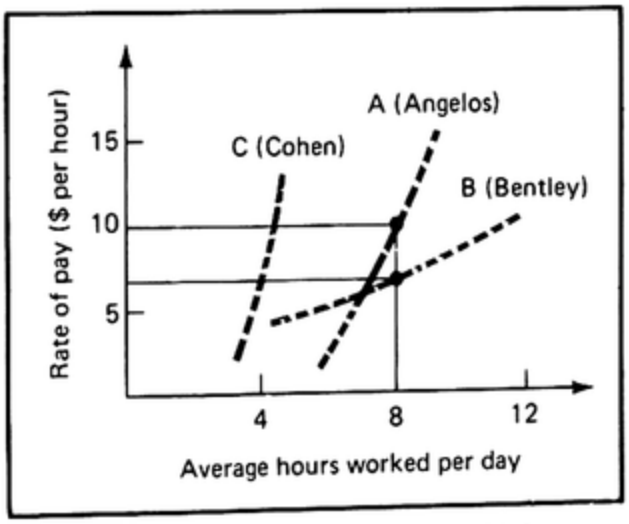


Figure 2 Individual supply curves of labor

The supply curves all slope up in this range of pay. Higher rewards induce longer hours, by offsetting the marginal disutility of work. Viewed differently, higher pay rates make the marginal hours of leisure more expensive because each extra hour of leisure results in more and more income lost that could have been earned.

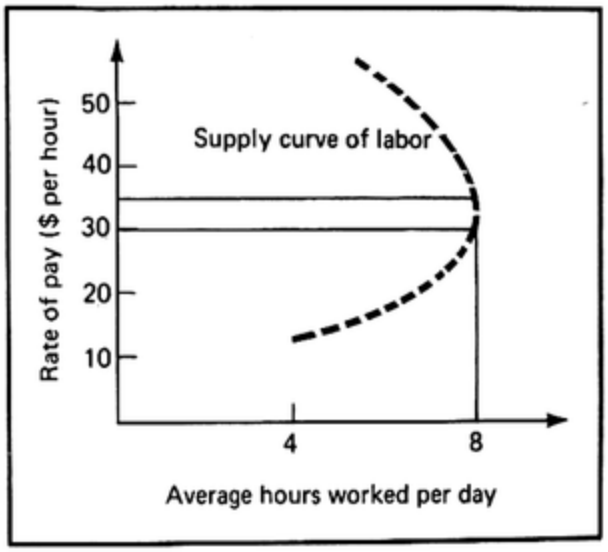


Figure 3 A backward-bending supply curve for labor

At higher rates of pay, a person's supply curve for labor may bend backward. The *price effect* increases work effort by increasing the opportunity cost of an hour of leisure. The *income effect* reduces work effort by increasing a person's ability to purchase leisure. If the income effect overcomes the price effect, higher pay will reduce the number of hours worked, and the person's labor supply curve will bend backward.

In this diagram, pay rate increases up to \$30 cause the person to work longer hours, and the price effect is dominant. Increases from \$30 to \$35 per hour result in constant work effort; the price and income effects just balance. Pay rates higher than \$35 per hour cause the person to work less, and the income effect is dominant.

Since workers' preferences and needs are so diverse, a range of jobs with flexible schedules is needed to match the variety of individual preferences.

Price and income effects

The labor supply curves in Figure 2 all slope upward. Yet, it is possible for labor supply curves to become vertical and then slope backward, as shown in Figure 3. As pay rises, people may choose to work *fewer* hours. The reason for this reversal of slope is the dual operation of a ***price effect*** and an ***income effect.*** As wages increase, the opportunity cost of an hour of leisure also increases. If wages are \$5 an hour, substituting an hour of leisure for an hour of

work means a sacrifice of only $5. If wages rise to $10 per hour, the opportunity cost of an hour of leisure would rise to $10.

The increasing cost of leisure resulting from higher wages is referred to as the *price effect*. The wage rate is, in fact, the price of leisure. As this price rises, people tend to work more and "consume" less leisure. This helps to explain the upward-sloping portion of the labor supply curve: Higher wages mean more work and less leisure because the cost of leisure is high.

However, the *income effect* also operates here. As wages rise, so does your income. With this rising income, you can afford to consume more of all goods, including leisure. For example, if you are paid $15 per hour for 8 hours of work per day, 5 days per week, 48 weeks per year, you earn $28,800 per year. A wage increase to $20 per hour would give you $38,400 a year, an increase of $9,600. If you feel that $30,000 is about all you really need, you might trade off some of the increased pay for increased leisure, by working only 7 ½ hours per day.

Thus, as the rate of pay rises, the price effect increases the cost of leisure, while the income effect increases one's ability to purchase leisure. If at some point the income effect overcomes the price effect, then a person works less for more money, and the supply curve bends back.

Virtually everyone's labor supply curve reaches a range of backward slope at sufficiently high wage rates. For most people, that range occurs well above the wage rates they can actually get for their skills. Therefore, only the positive-sloped portion is relevant for their actual choices.

The market supply curve of labor

To obtain the market supply curve of labor, all persons' labor supply curves are added horizontally. Figure 4 shows how the summation is done. The backward-bending parts of most people's curves occur at high rates of pay, well above the prevailing wage rates. Since most people are responding to wages on the upward-sloping portion of their supply curves, the market *summation* of individual curves usually slopes up rather than back. In Figure 4, for example, the market supply curve still slopes up, even though Person B's supply curve bends backward at wages above $12 per hour.

The market supply for specific kinds of labor will slope up, not only because of the shape of the individual supply curves, but also because new people will be attracted to work in the industry as the wage rate rises. If the pay for secretaries is increased, other things being equal, people who are already secretaries may work more. Salespeople and clerks—or even teachers or pipefitters—may also decide to become secretaries. Moreover, the higher pay can attract workers from other locations. Rising wages in Houston, for example, have drawn new workers from as far away as Michigan and Maine. The upward slope, then, represents both longer hours and more employees.

The balance between work and leisure, decided by millions of workers, yields the overall pattern of participation in the entire labor force. Such overall patterns are called *labor force participation rates*. They strongly affect the operation of the entire economic system.

Figure 5 shows the participation rates for people of various age groups and marital status. The participation rate is the share (up to 100 percent) of each group that works for pay. On the horizontal axis, the main working years are about ages 20 to 65.

Over 90 percent of the men between ages 20 and 65 work for pay. Whether women work depends crucially on their marital status. Most unmarried women (single, divorced, or widowed) have paid jobs. About 40 percent of married women

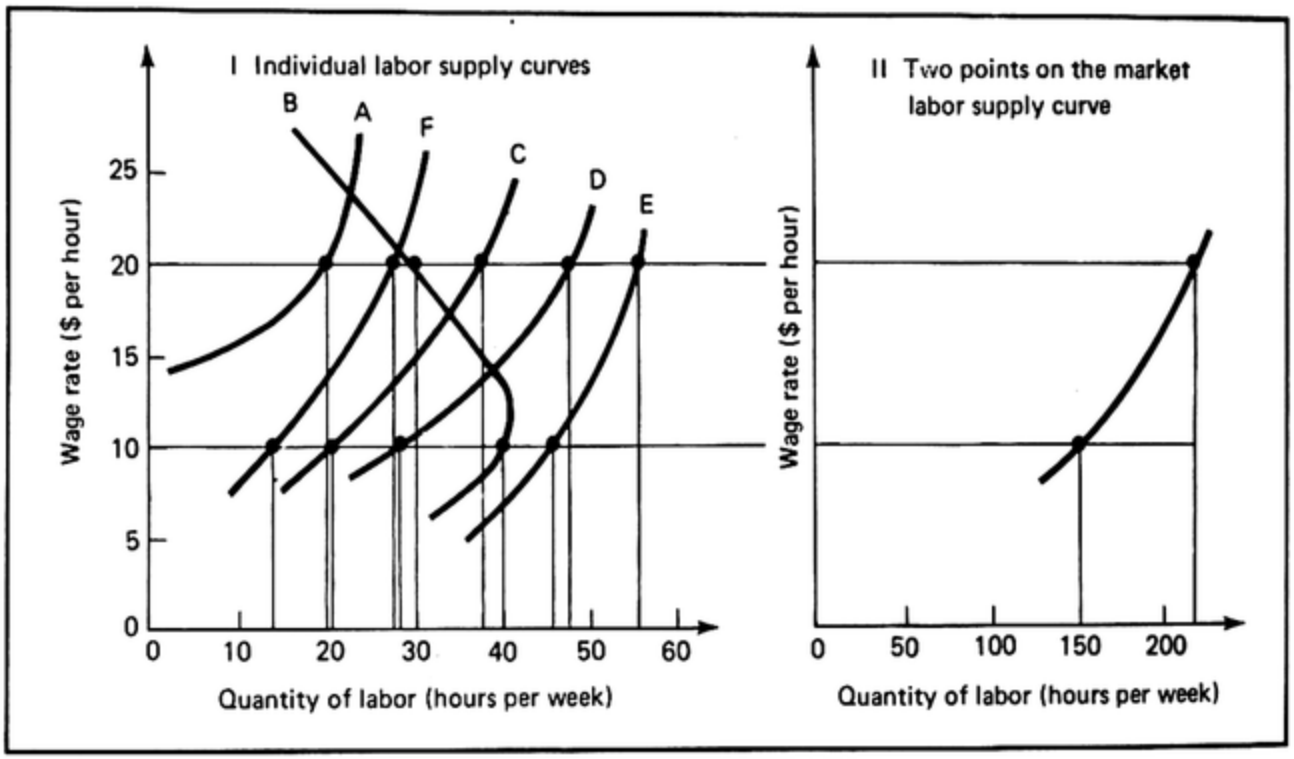
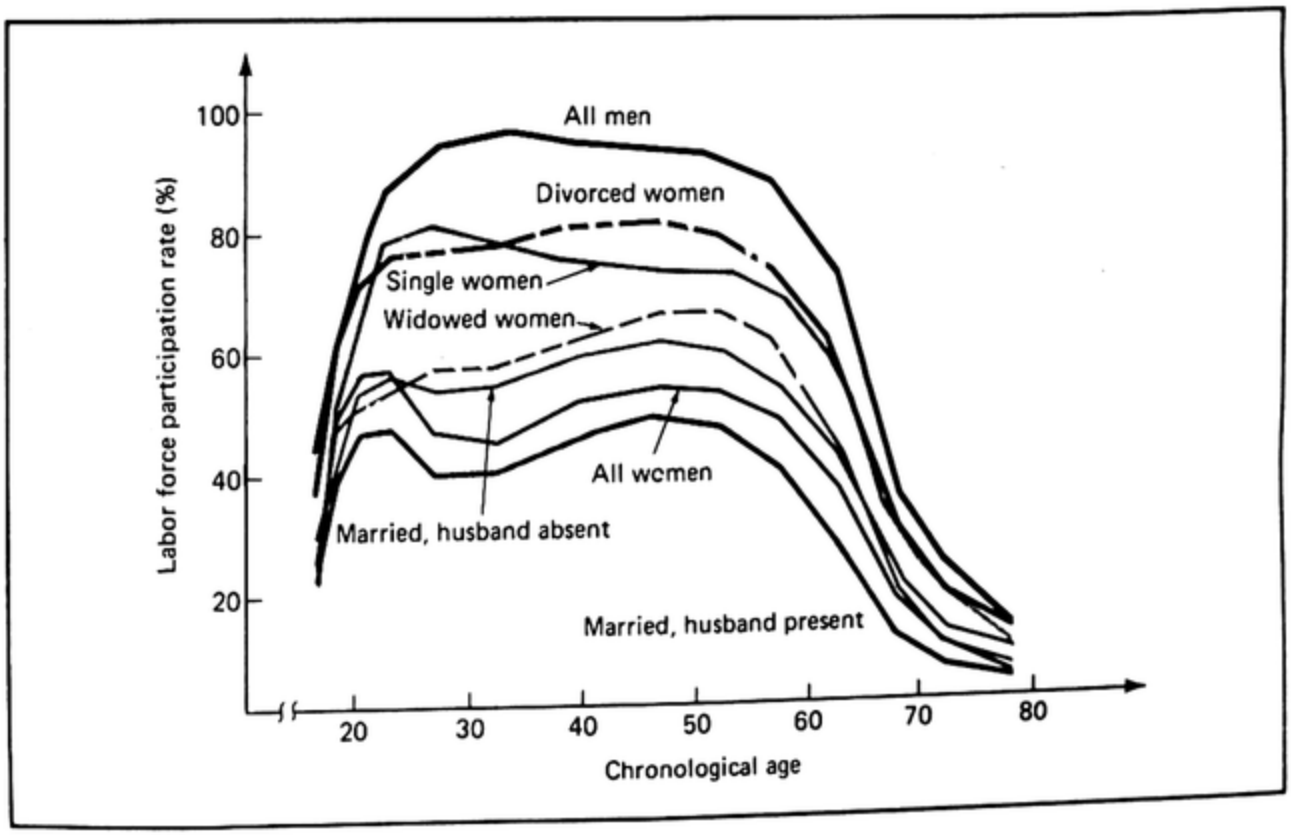


Figure 4 Summing up individual labor supply curves to obtain the market supply curve

Labor markets usually consist of hundreds or thousands of people. Here, the labor supply curves of only six people are summed to show how the *market* supply curve for labor is derived. At a pay rate of $10 per hour, the six people will work for a combined total of 148 hours. At $20 per hour, the same six people are willing to supply 220 hours of labor. By adding the combined labor-hours these people will offer at various wage rates, the entire market supply curve can be obtained.

Figure 5 Labor force participation rates by age for men and women by marital status, 1970



work outside the home, but fewer work during the child-rearing years of 25 to 40. The average rate for all women has risen from about 50 percent in 1970 to about 60 percent in 1980.

The overall supply of labor to the economy involves the decisions of a complex variety of people in many circumstances. Their decisions about work are not made unilaterally, however. What people can actually achieve in terms of hours of work and pay rates depends not only on their own preferences, but also on the demand for labor.

The demand for labor

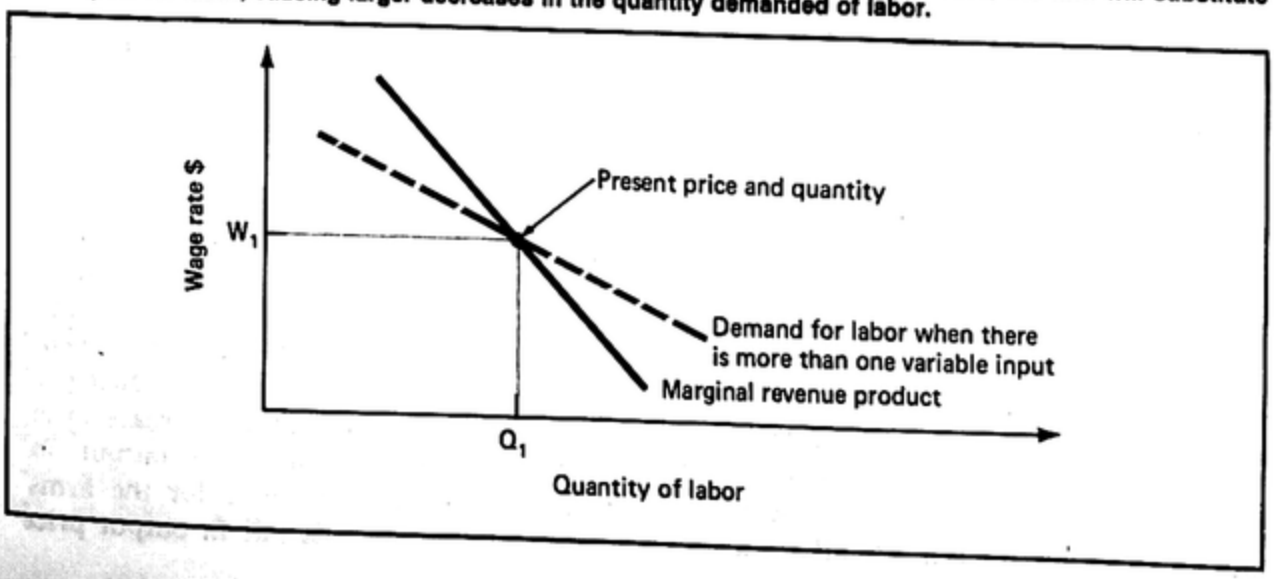
The demand for labor is, like the demand for any other input, a *derived demand.* It derives from the demand for the output that the input helps to produce. Firms will hire workers only up to the point at which the value added by the workers to output is equal to their wages cost. As with other inputs, the firm's demand curve for labor is the *marginal revenue product curve: the amount of output that one additional unit of labor can produce (the MP) times the additional revenue from the additional output (the MR).**

The *height* of the firm's MRP curve for labor, in Figure 6, depends mainly on two conditions: (1) the amount of other inputs used; and (2) the price of the output. If the amount of other inputs is increased—for example, by giving the workers bigger and better machines to work with—then the marginal product curve of labor shifts up, and so must the marginal *revenue* product curve. If the market demand for output increases, so that the marginal revenue curve for the output shifts up, the mar-

*Remember that the MRP curve represents the demand curve for an input if the firm is a profit maximizer, a price taker in the input market, and uses one variable input.

Figure 6 Contrast of a firm's demand schedule for labor with one variable input and with several variable inputs

For one variable input, the marginal revenue product schedule is the demand schedule for labor. For more than one variable input, the demand schedule for labor will be more elastic than the marginal revenue product schedule because substitution among the variable inputs will cause a larger change in the quantity demanded of inputs in response to changes in input prices. As wages fall below W_1, the firm will substitute labor for other inputs, causing larger increases in the quantity demanded of labor. As wages increase, the firm will substitute other inputs for labor, causing larger decreases in the quantity demanded of labor.



ginal revenue product curve must also shift up and to the right. Finally, the downward *slope* of the marginal revenue product curve reflects the law of diminishing marginal product.

Remember, though, that the firm's demand curve for labor is the marginal revenue product curve only for cases in which labor is the only variable input. It is interesting to see what happens to the demand for labor when there is more than one variable input.

Figure 6 shows the firm's marginal revenue product schedule, the demand schedule for labor if there is one variable input. The curve slopes downward, indicating that the firm will hire more labor as wages decrease. The reason for this, as you saw in Chapter 14, is that as the price of a variable input falls, the firm's marginal cost schedule shifts down and to the right. This causes the profit-maximizing level of output—the point at which marginal cost equals marginal revenue—to increase. To produce more output, the firm must hire or buy more inputs.

When there is more than one variable input, there is another reason for the firm's demand for labor to change: substitution among inputs. As wages fall, for example, labor becomes relatively cheaper compared to other inputs. The firm would, therefore, try to substitute labor for its other variable inputs in the short run. When wages fall, the firm will hire more labor, *both* to produce more output *and* to substitute for other inputs. The total change in demand for labor, then, will be greater in the case of several variable inputs, assuming that the firm's technology permits substitution among inputs. For decreases in the cost of labor, the firm's demand for labor will lie to the right of its marginal revenue product schedule.

If wages rise, the firm's marginal cost schedule will shift to the left. The profit-maximizing level of output falls and will cause the firm's demand for labor to fall. If the firm uses more than one variable input, the firm will also attempt to substitute other inputs for the relatively more costly labor. Again, the total change in demand for labor will be greater in the case of several variable inputs, assuming that technology permits substitution among the inputs. For increase in wages, then, the firm's demand curve for labor will lie to the left of the marginal revenue product curve, indicating larger decreases in the quantity demanded of labor.

Since both increases and decreases in wages will lead to larger changes in the quantity demanded of labor when there is more than one variable input, the resulting demand schedule will be more elastic in the case of several variable inputs. The degree of elasticity will depend on the ease of technical substitution. The easier the substitution, the more elastic the demand for labor will be, all other conditions being equal. This contrast between the marginal revenue product schedule as the demand schedule for the input in the case of one variable input and the more elastic demand schedule for labor in the case of several variable inputs is shown in Figure 6.

In some cases, a market supply or demand schedule can be derived by summing up individual supply or demand schedules. For labor, however, the derivation of the market demand schedule for labor through a summation of the firms' demand schedules for labor is not strictly accurate. The reason is simply that changes in labor costs may cause industry output and, therefore, prices to change, resulting in shifts in the firms' labor demand schedules. For example, suppose that labor costs fall. This will cause the marginal cost schedules of all firms in the industry to shift down, resulting in an increase in industry output. As industry output increases, the industry price for the firms' output will fall. This fall in output price

will cause the firms' demand schedules for output and, therefore, the marginal revenue schedules to shift down. As the marginal revenue schedules shift, the firms' demand schedules for labor will also shift. If a change in labor costs can cause the firms' demand schedule for labor to shift, a summing up of existing labor demand schedules will not accurately reflect the total quantities of labor that the firms in the industry will use at different prices.

While the market demand schedule for labor cannot be derived from a horizontal summation of the firms' demand schedules for labor, the factors that will influence the firms' demand for labor will obviously also influence the market demand for labor. For example, increases in labor productivity or in the demand for the firms' output will shift the firms' demand schedules for labor to the right, and also shift the market demand for labor. Increases in the number of firms operating in the industry will also cause a rightward shift in the market demand schedule for labor.

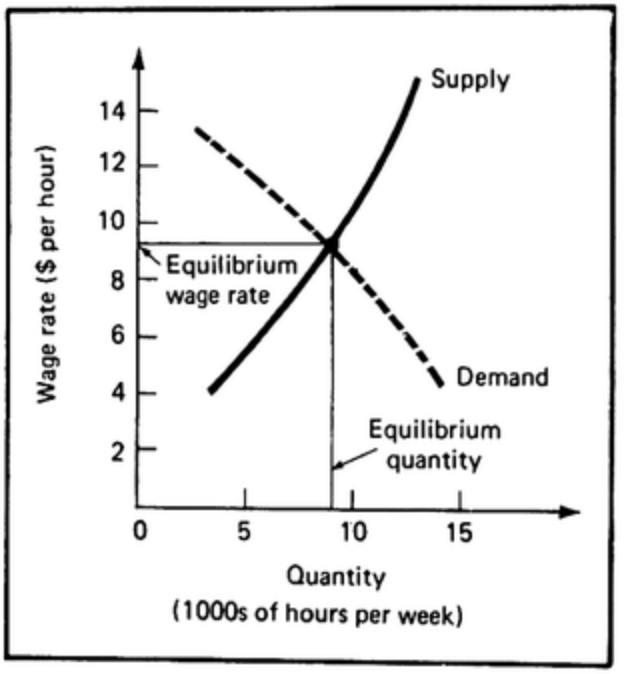


Figure 7 The demand for and supply of labor in each market determine the equilibrium wage rate and the amount of labor hired

Market equilibrium is reached at the point where the market clears, with the quantity supplied of labor just equal to the quantity demanded. In this diagram, equilibrium is reached at a wage of $9.25 per hour with 9,000 labor hours purchased. At higher wages, the excess supply of labor would bid the wage down. At wages lower than $9.25, the excess demand for labor would cause the wage to rise.

Equilibrium between supply and demand

Once the market supply and demand curves for labor are derived, the equilibrium quantity and wage in that labor market can be determined. The intersection of supply and demand is the equilibrium point at which the market clears. In Figure 7, for example, the market equilibrium occurs at a wage of $9.25 and quantity of 9,000 labor-hours. The equilibrium quantity of 9,000 hours of labor per week reflects both the number of people working and the average number of hours that they each work. At wages higher than $9.25, there would be an excess supply of labor, with unemployed workers bidding the wage level down. At wages below $9.25, there would be an excess demand for labor, with eager employers bidding wages up. In both cases, the wages would be pulled back toward the equilibrium wage of $9.25.

The important distinction between changes in supply and demand and changes in quantity supplied and quantity demanded still holds. A change in wage levels (the price of labor) will cause changes in quantity supplied and quantity demanded, shown by a movement along the supply or demand schedules for labor. A change in supply or demand refers to a *shift* in the entire supply or demand curve. A change in technology or in demand for output would cause a shift in the *demand* schedule for labor. A change in people's work preferences or in the size of the labor force would cause a change in the amount of labor offered at every wage, shifting the labor *supply* schedule. Some possible shifts

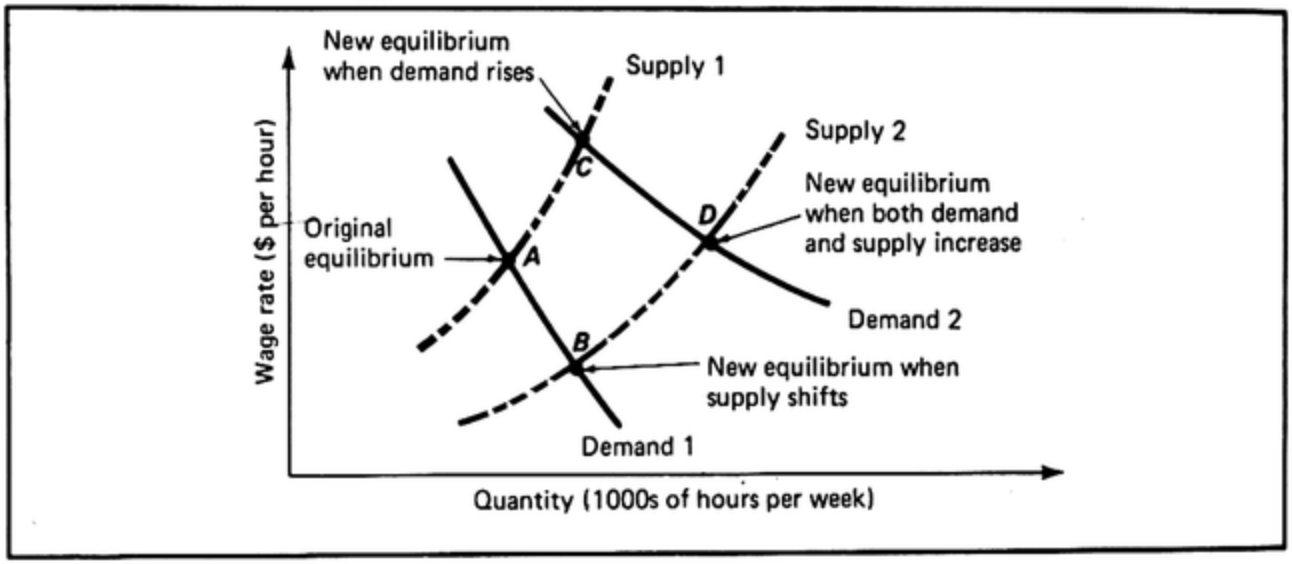


Figure 8 Shifts in the demand for and supply of labor

If the *supply* of labor increases, the supply curve will shift to the right from Supply 1 to Supply 2. The new equilibrium at Point *B* will show a lower wage rate and a higher quantity of labor hired than the original equilibrium, represented by Point *A*.

If the *demand* for labor increases, the demand curve will shift to the right, from Demand 1 to Demand 2. The new equilibrium (Point *C*) will show a higher wage and higher quantity of labor hired than the original equilibrium.

If both the supply of and demand for labor increase, the equilibrium quantity of labor hired will increase, but the net effect on the wage rate is uncertain, depending on the relative size of the shifts. The wage rate will fall if the supply shift dominates and rise if the demand shift dominates. In this diagram, the demand shift dominates, and there is a slight increase in the wage rate.

and their impact on the labor market's equilibrium outcomes are shown in Figure 8.

Like other supply and demand schedules, the labor supply and demand schedules have elasticities that reflect underlying market conditions. The elasticity of demand for labor will depend on both the elasticity of demand for output and the importance of labor costs as a percentage of total costs. As Chapter 14 showed, the elasticity of demand for an input such as labor will be greater the more elastic is the demand for output and the larger is the percentage of total costs for which labor accounts.

The elasticity of the supply of labor will depend on both the responsiveness of present workers to wage increases and the ease with which new workers can be attracted to the particular labor market. The more willing present workers are to increase the amount of work they do as wages increases, the more elastic the supply of labor will be. Moreover, the easier it is for new workers to enter the job market, the greater the elasticity of supply will be. For example, consider the supply of neurosurgeons. Because access to a medical school education is limited, the increase in the supply of doctors in response to a wage increase is restricted. Training to become a neurosurgeon is a lengthy and costly process, and not many people have the talent for it. All of these factors tend to make the supply of neurosurgeons relatively inelastic. This is particularly true for short-run elasticities because workers cannot easily shift into this profession from other fields.

In contrast, the supply schedule for part-time waiters is relatively elastic. The flexibility of hours would attract many people; the training period is short and not costly to the worker. Since the specific skills required to wait on tables are not numerous, there can be shifts to such a job from other occupations, like clerking in a store, or simply from new additions to the labor force. The nature of the job, then, in terms of the skills required and the length

and cost of training, is a major determinant of the elasticity of supply.

Differences in labor skills

The actual variety of jobs in any modern economic system is staggering. The standard system of job classifications divides jobs into ten main classes, as shown in Table 1. Each of these main classes contains a great diversity of jobs in terms of specific skills, physical and mental effort, and job conditions. There are ***blue-collar jobs*** that deal directly with production, such as drill press operators and farmers. ***White-collar work*** is done mainly in offices.

Variations in pay rates

With these job variations in mind, consider the basic wage patterns and trends in the U.S. economy, as presented in Table 2. Though white-collar jobs pay better than blue-collar work on the whole, skilled

Table 1 Varieties of jobs

Category	Specific instances
	White-collar workers
Professional, technical, and kindred workers	Physicians, dentists, editors, osteopaths, engineers, lawyers, chemists, teachers, pilots, architects, accountants, etc.
Managers and administrators, except farm	Buyers, purchasing agents, sales managers, school administrators, public administrators, bar managers
Sales workers	Insurance agents, brokers, and underwriters, real estate agents and brokers, sales clerks in retail stores, peddlers
Clerical and kindred workers	Bank tellers and cashiers, bookkeepers, billing clerks, mail handlers, postal clerks, file clerks, typists, etc.
	Blue-collar workers
Craftsmen and kindred workers	Electricians, plumbers, automobile mechanics, TV repair workers, carpenters, bakers, sheetmetal workers, printers, upholsterers, etc.
Operatives, except transport	Meat cutters and butchers, gas station attendants, laundry workers, welders, packers and wrappers (except product), etc.
Transport equipment	Truck drivers, delivery workers, taxi drivers, bus drivers, chauffeurs
Laborers, except farm	Construction laborers, freight, stock, and material handlers
Farmers, farm managers, and farm laborers	
Service workers, except private households	Janitors, cooks, waiters, waiters' assistants, dishwashers, fire fighters, security guards, nurses' aides, orderlies, hairdressers, policemen, etc.

blue-collar workers earn more, on average, than salespeople and clerks. In fact, a good plumber may earn more than the average college professor. One of the most noticeable features of the table is the impact of skills on pay rates. The pay gap between unskilled and skilled work, for both white-collar and blue-collar jobs, is large.

Around the average earning figures, shown in Table 2, lie infinite variations in earnings, caused by a particular worker's age, experience, specific company, region of the country, and, of course, productivity. Although many factors influence pay for both individuals and different lines of work, two main causes of wage differentials are the costs of training and the scarcity of talent. These factors are examined in this section of the chapter. Monopoly power of either the buyer or seller of labor can also influence levels of power. This issue is examined in the third main section of this chapter.

Investment in human capital: The cost of training

Labor is not just the use of raw, muscular force. Labor is, in fact, the use of ***human capital,*** the talents and skills created by

Table 2 ***Broad and specific variations in incomes***

Category	Median yearly earnings, full-time workers, 1980
White-collar	
Professional and technical workers	$19,656
Managers and officials	22,100
Clerical workers	11,960
Sales workers	14,664
Blue-collar	
Craftsmen and foremen	$17,368
Operatives	12,480
Nonfarm laborers	9,100
Private household workers	3,848
Other service workers	9,360
Farm workers	7,696
Specific occupations	
Physicians and surgeons	$82,000
Executives	76,000
Lawyers	47,000
Dentists	46,000
Airline pilots	44,000
Electricians	34,000
Economists	31,000
Automobile workers	24,900
Accountants	24,000
Insurance salespeople	21,000
High school teachers	20,000
Computer programmers	18,000
Typists	11,000
Retail salesclerks	10,500
Unskilled labor	7,400

Source: U.S. Labor Department and Census data, adjusted from 1970 information; *Medical Economics* (for physicians and surgeons); and various other sources. Most figures are approximate and do not include nonsalary benefits.

Investing in College

You are a costly bundle of human capital already, on the way to becoming even more valuable. Consider the average child of a middle-income family who goes to a four-year public college. The average family's direct costs for that average child are $85,000, in 1980 prices. The table shows the detailed estimates from a recent study.

Besides the direct family investment there are other costs. Public schooling from kindergarten through senior year of high school costs about $1,000 per year, or about $13,000. Moreover, the earnings forgone while at college may be about $7,000 per year, or $28,000 in total. On top of all these expenses are the lost earnings of parents who choose to stay home with their children. For example, for a parent who stays home until the child is ten years old, giving up a $10,000 yearly job, the opportunity cost is $100,000. (Of course, if both parents work, day-care expenses must be calculated into direct costs.)

All of these direct and indirect costs add up to an average expenditure of about $230,000 by the time the student graduates from college. Studies after college or on-the-job training and experience will increase the total amount of that investment even more.

Average investment in a child by graduation from college

Category	Amount (in 1980 prices)
1. *Direct family costs*	
Housing	$24,711
Food	17,931
Transportation	12,027
Public college for 4 years	9,784
Clothing	5,686
Medical	3,716
Childbirth	2,485
Educational materials	1,020
All other expenses	7,726
Total	$85,086
2. *Public schools cost* (13 years at $1,500 per year)	19,500
3. *Earnings forgone by student while at college* (4 years at $7,000 per year)	28,000
4. *Earnings forgone by parents while raising child* (10 years at $10,000 per year)	100,000
Total	$232,586

Source: Family costs from Thomas B. Espenshade, *Raising a Child Can Now Cost $85,000*, INTERCOM, 8, no. 9 (Washington, D.C.: Population Reference Bureau, September 1980).

investment in people. Each person is born with certain unique qualities, or talents. Talent is a raw material that is made more productive by education and training. Each person is like a complex machine, with productive skills built up by a long process of economic investment. The investing begins at birth, continues with feeding and sheltering, then expands with schooling and specialized training.

The economic theory of human capital suggests that people will invest in schooling and/or other types of training up to the point where the marginal return to training, measured in terms of future income gains, equals the marginal cost of training. Those costs include both the direct costs (the cost of tuition at a college or university, for example) and indirect costs, such as the income given up to receive more schooling. This theory suggests that talented people will invest more in education, since their marginal returns are likely to be higher; it also predicts that those who have access to cheaper funding (because of affluent family backgrounds, for example) will invest more in training.

These predictions generated by the theory of human capital are confirmed by actual data. The effect of educational investment on later earnings is shown in Figure 9. Workers with a high school education had earnings well above those who had only completed the eighth grade. College graduates' earnings were much higher still.

The extra earnings are a form of return from the added educational investment in human capital. Economists have investigated the rates of return on this investment. A representative recent set of results is given in Table 3. These results fit the predictions of economic theory. Grade schooling provides rates of return above 20 percent. Higher grades provide decreasing marginal returns, but postcollege work still produces an average return on investment of over 7 percent. This is comparable to the returns earned by most other forms of capital. Therefore, the process of education and training resembles other kinds of productive investment.

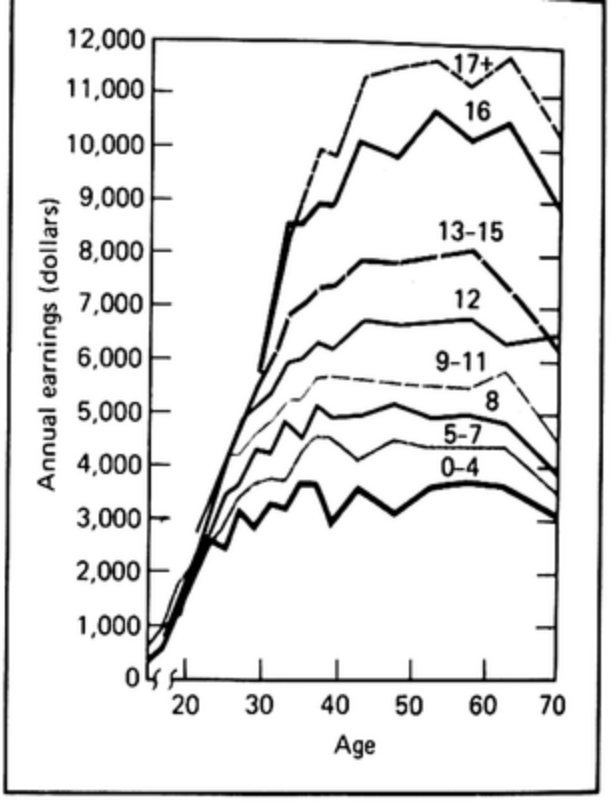


Figure 9 Age profiles of earnings of white, nonfarm men 1959 (annual earnings classified by years of age, for indicated schooling groups)

NOTE: Figures on curves indicate years of schooling completed.

Source: Jacob Mincer, *Schooling, Experience and Earnings* (New York: National Bureau of Economic Research, 1974), p. 66.

Table 3 Estimated rates of return to investment in human capital: Various levels of education

Additional year of schooling	Private rate of return in the form of higher income
8th year	22%
11th year	16%
12th year (senior year of high school)	16%
16th year (senior year of college)	12%
17th and higher years	7%

Source: J. T. Addison and W. S. Siebert, *The Market for Labor: An Analytical Treatment*. Copyright © 1979, Scott Foresman and Company. Reprinted by permission.

To view people as the embodiment of valuable investment does not demean them. The analysis of human capital merely recognizes the economic truth that skills are costly to develop. A skilled population is a valuable resource. The productivity of labor—for each person and for all of us collectively—reflects the investments made in human capital, both today and in the past.

Since investment in human capital is costly, the pay that workers receive must compensate them in the long run for the training that they acquire. If the benefits of an occupation do *not* equal the costs of preparing for it, the supply of workers will soon decrease. Consider a specialized doctor who has trained until age 35 before opening a medical practice. Assume that the doctor's true costs of postcollege training (including forgone earnings) were $240,000. Between age 35 and retirement at 65, the doctor would need to earn at least $240,000/30 = $8,000 more each year than other college graduates, plus interest on the investment, to offset those additional training costs. After all, many people are unwilling to undertake additional training now unless they feel that it will "pay off" in higher income later. Therefore, the higher income can be interpreted as a necessary incentive to induce candidates to undertake training.

The doctor may, in fact, earn much much more than that extra $8,000 a year. If, instead, he or she earns $50,000 extra per year, only a small fraction of it would be explained by the cost of training. Other factors must be at work.

Scarcity of talent

Scarcity is a theme that underlies most economic analysis, and wage determination is no exception. For any given level of labor quality—involving skill, strength, and creativity—greater scarcity, represented by the height of the supply curve, will result in higher pay. Increasing scarcity results in a leftward shift of the supply curve, which will cause the equilibrium wage to increase. How much wages will rise depends on the elasticities of labor supply and demand. The more inelastic supply and demand are, the larger is the wage increase that an increasing scarcity of labor will bring. A greater elasticity of supply and demand will result in smaller wage increases as labor grows relatively scarce.

The influence of labor scarcity can be seen in any number of everyday situations. Picking tomatoes requires effort and endurance, but because the supply of tomato pickers is ample relative to demand, wages are low. The talent of a John McEnroe, a Bruce Springsteen, or a Luciano Pavarotti is rare, and these performers earn enormous salaries. If there were 200 people with talents identical to Bruce Springsteen's or Luciano Pavarotti's, these men might earn much lower wages.

There is a *natural scarcity* of many talents, although training and hard work can enhance average abilities. *Artificial scarcity* can also occur, when people or organizations deliberately exclude qualified or potentially qualified people from a certain job market. The medical profession is often accused of doing this, by limiting medical school enrollments to hold down the number of doctors. Some 50 professions—from the law and dentistry to hairdressing and undertaking—are "protected" by state licensing laws, which help to limit the number of people in these professions. Of course, licensing restrictions do more than keep the supply of labor low. Few consumers would want to be attended by incompetent dentists or doctors.

You can see that many factors such as productivity, training costs, workers' preferences, demand for output, and scarcity interact to determine prevailing wage

rates in each job market. As individuals work their way through the educational process, they consider various job possibilities in the light of their own skills, the possibilities for enhancing present skills or acquiring new ones, and differences in wage rates. On the basis of these factors, they begin to make job-related choices. If the supply of labor in a particular labor market is high relative to demand, wages in that market will fall. Fewer people will be attracted to that line of work, and the supply and demand imbalance will disappear. If the demand for labor in a certain job market is high relative to supply, wages will rise. New entrants will be attracted to that job market, and the shortage of labor will correct itself.

Up to this point, the discussion of labor markets has focused on situations in which no group of suppliers or demanders of labor—neither the employees nor employers—controls the labor market. Yet there are important cases in which this assumption does not hold. Workers (the suppliers of labor) can join together to form monopolies such as unions on the supply side of labor markets. Buyers of labor services may hire such a large percentage of the workers in a given labor market that the firm becomes a monopoly on the buying side, which economists call a *monopsony*. These two types of departures from the competitive functioning of labor markets are examined in the next main section.

Departures from competitive market outcomes

Labor markets usually include hundreds or thousands of workers, making these markets inherently competitive on the supply side. No one worker can exert any market power. Yet, if workers can form ***unions*** or other associations, they may be able to control some of the supply side of the labor market. The group can then behave like any other monopoly, limiting the quantity of labor supplied while raising its price or wage rate. The labor monopoly may also seek other goals, such as improved work conditions, and greater stability of employment.

To maximize its control, the union commonly seeks to unify all the local unions into one industry-wide union. The union can then apply the strategy that works best: dealing with the weakest firms first, for example.

Barring entry by nonmembers can be crucial. Each worker has conflicting motives toward the union. The worker can gain increased wages by joining and supporting the union. Yet, it is also possible to gain from having a successful union but *not* joining it. Such "free riders" get the higher wages but don't have to pay the union dues or sacrifice their incomes by supporting a strike. Unions at one time tried to negotiate a *closed shop*, with only union members eligible for hiring. Closed shops have been illegal, however, since the Taft-Hartley Act of 1947. Another approach is the *union shop*. In this case, nonunion members may be hired if they agree to join the union within a specified time (union shops are illegal in 20 states).

If the union fails to limit the labor-market entry of nonmembers, support for the union dwindles. Each worker relies on the *others* to support the union. The company, of course, would like to have no union at all. Failing that, it would prefer its own company union or an *open shop* in which union membership is not required for employment. Unions have less control over the labor supply in an open shop and therefore tend to have much weaker bargaining strength than unions that can restrict the hiring of nonmembers.

Labor's control over a particular market is reflected in the elasticity of demand

for the union's labor. An industry-wide closed-shop union would have a relatively inelastic demand for its members, since the industry has no legally available substitutes for union labor. That gives it more ability to raise wages by restricting the quantity of labor supplied. In contrast, an open-shop union would face a more elastic demand because its members can be replaced by the hiring of nonmembers. The high elasticity means that wage increases will result in a relatively large decrease in the number of union members employed.

Leverage: The power to inflict economic damage Once a union controls the labor supply in a particular market, the extent of its power to extract better pay and work conditions depends largely on its ability to inflict economic damage on the employer. This damage occurs as work stoppages or *strikes*. A successful strike stops production. The struck firm loses sales and, therefore, profits, perhaps experiencing large financial losses. Customers of the struck firm lose, too, especially if it produces a crucial good or service. The suppliers of inputs to the struck firm also lose sales, and their profits decrease. The disruption often spreads further. Stores that sell to the workers suffer, since striking workers have less money to spend (even though they usually receive some strike pay from the union) and therefore buy less. Since local government tax revenues are cut along with the strikers' income, local services can also be affected. Disruption from a large-scale strike can be severe and widespread.

The greater the possible disruption from a strike, the more seriously an employer will view a strike threat. The larger, therefore, will be the gains that the union can force the employer to yield. To maximize the threat of a strike, the union will choose the most favorable time to strike, the time that will most adversely affect a firm's production. The very threat of a strike may be sufficient to gain the union's demands, or the employer may refuse to give in, and the strike may occur. Whether the strike occurs depends largely on the employer's weighing of the costs of the strike against the costs of giving in to the union's demands.

Employers' losses from a strike depend on the ease with which striking workers can be replaced. This depends on several factors. First is the specific role the striking workers play in the technology of production. Highly skilled workers, such as operators of giant oil refinery equipment and hospital staff doctors, play crucial roles. By contrast, unskilled or low-skilled workers, such as tomato pickers and janitors, are generally easy to replace because there is a large and readily tapped reservoir of available workers. Thus, unskilled workers rarely win important strikes.

Second is the effective cohesion of labor. If labor sympathy for the striking workers is strong, or if the union can enforce a complete shutdown and block replacement workers, the union's leverage will increase. Finally, the union's leverage is affected by the ability of firms to relocate production, both immediately and in the long run. "Multinational" firms that have parallel production facilities abroad can often break a strike merely by shifting their production to a foreign plant. A nationwide firm can shift production to new plants in less unionized parts of the country. In general, the ability of firms to threaten job losses by shutting down plants and permanently relocating production can eliminate a union's leverage.

Yet, employers and unions must consider not only how a strike will damage them, but also how it will hurt others. A doctors' strike threatens the sick. A strike against a canning firm harms vegetable growers, whose crops may rot in the fields.

A telephone strike stops communications throughout the nation, affecting nearly everyone. All of the groups harmed by a strike will pressure the employers to come to terms.

In general, the union's leverage and the likelihood that it will prevail increases with the amount of damage that a threatened strike poses to all the parties that might be affected by it. When the threat of damage becomes especially severe, however, as with special groups such as doctors, nurses, police, teachers, and fire fighters, the groups are often forbidden to strike by law. Or society believes that they have a duty not to endanger others by striking. The group members themselves may consider strikes unethical or "unworthy" of them.

The choice, then, of whether to use the strike as a bargaining tactic depends both on the calculations of the economic damage of the strike and on the labor organization's view about whether the strike is an appropriate bargaining tool.

The choice of goals and methods Given whatever degree of control and leverage it has, the labor group must choose its economic goals and tactics. The demand curve for its labor is down-sloping. The union can obtain higher wages for its members, but doing so will reduce the number of workers hired. The union's desire to raise wage rates will clash with its dislike of cutting the numbers hired.

In this dilemma, the elasticity of the demand curve for labor is crucial. If the demand curve is highly inelastic, wages can be forced up to high levels without reducing the number of jobs by very much. A highly elastic demand curve will, by contrast, leave little room for wage rises; even a small increase may drastically reduce the number of workers hired.

Most unions are not just profit-maximizing monopolies. They have a range of other goals, and their costs are not clearly defined. Therefore, economists usually set aside the profit-maximizing approach and, instead, compare the likely results of unions with the competitive market outcome.

To clarify unions' choices, we divide the discussion into two parts, reflecting two kinds of unions and their characteristic strategies: (1) *industrial unions*, which try to enroll all workers and then set higher wages for them; and (2) *craft unions and professional organizations*, which mainly attempt to reduce the supply of labor in their trades.

The typical *industrial union's* choices are shown in Panel I of Figure 10. In this case, the wage that will maximize the total payment to labor is higher than the competitive equilibrium wage. In Panel I, the levels W_1 and Q_1 represent the competitive equilibrium determined by the intersection of the supply and demand curves at Point *A*.

To determine the wage that will maximize the wage bill, one needs to examine marginal revenue. As long as the addition to labor revenue (which is the union group's marginal revenue) is positive, the wage bill increases as employment rises. Thus, just as with sales revenue, *total labor revenue* (W × Q) *is also maximized where marginal revenue is zero.* In Figure 10, the wage bill will be maximized at a wage of W_2 with Q_2 workers hired. The result is clearly a wage higher than the competitive level, but with fewer workers hired. As Panel I shows, employment will decrease from Q_1 to Q_2 workers. The total amount of dollars paid to workers is now the rectangle OW_2 BQ_2, an increase from the competitive outcome.

The precise effects depend on the elasticities of both demand and supply. Inelastic demand (e.g., for "crucial" workers, who cannot easily be cut back) will result in a sharp wage boost for union labor, as

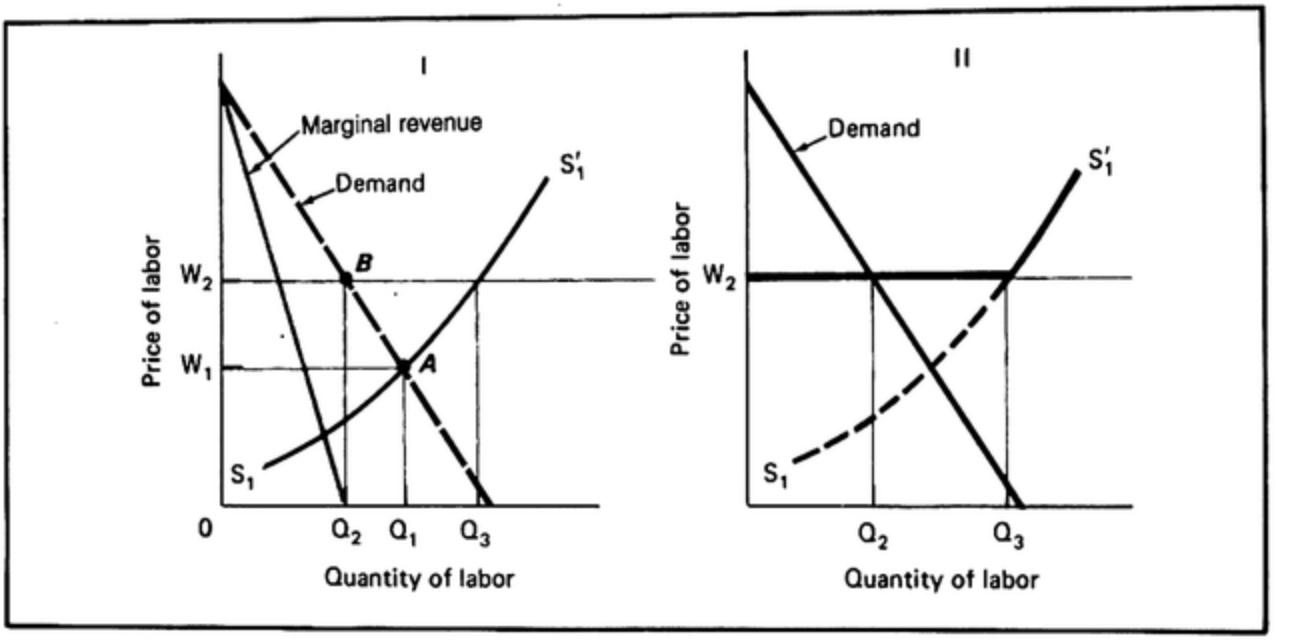


Figure 10 The possible effects of unions on wages and employment

In Panel I, competitive supply and demand conditions would result in an equilibrium wage of W_1 with Q_1 workers hired. The union, however, may try to maximize the total wage bill by bargaining for W_2 with Q_2 workers hired. The wage bill B (OW_2BQ_2) is larger than at Point A (OW_1AQ_1).

Panel II shows the supply curve perceived by the firm if the union negotiates a wage of W_2. The firm may hire as many workers as it wishes at the wage of W_2, up to a quantity of Q_3 workers. The supply curve of labor to the firm, therefore, becomes a horizontal line at the level of W_2. The actual number of workers hired will be determined by the firm's demand curve for labor. To obtain more than Q_3 workers, the firm must offer higher wages.

was noted above. At any rate, some workers will be unemployed at the new wage rate. Those not employed in that industry will eventually be absorbed elsewhere. Still, some workers (the amount Q_3 minus Q_2) will seek jobs here and not get them.

Of course, at a wage of W_2, more workers want the job than at W_1. The amount Q_3 shows how many would work at the wage of W_2, but only Q_2 will actually be hired. Since the union controls the supply of workers, however, the firm no longer perceives its supply curve for labor as S_1 S_1' as shown in Panel I. Rather, once the union negotiates a wage of W_2, the firm sees its labor supply curve as a horizontal line at the W_2 level, as shown in Panel II. Up to Q_3, the negotiated wage is higher than the supply curve, so that the firm can hire as many employees as it wants at the wage of W_2. To reach levels above Q_3, the firm will have to offer higher wages to attract more workers.

Alternatively, the union may seek to set even higher wages in the interests of a smaller group of workers—perhaps those with seniority. Total incomes will be smaller, but the remaining workers will have higher wages. Or the union may aim for a wage rate below W_2, to give wage gains to as many workers as possible. Typically, a union will not have a single precise goal. Rather, it will grope, estimating and compromising its objectives as it goes, in an uncertain environment. Yet, the likely range of those choices can be shown conceptually, as in Panel I of Figure 10.

Craft unions and professional groups Unlike industrial unions, craft unions and ***professional groups*** do not usually have to rely on the threat of strikes and similar actions

to achieve their goals. Craft and professional workers—plumbers, doctors, morticians, lawyers, electricians—usually sell their services to a mass of small-scale personal buyers. It would be inefficient to try to negotiate with so many employers. Instead, these service workers try to control labor market conditions by restricting or reducing the number of workers who can enter their fields. Craft unions require apprenticeship training before membership, and then tightly limit the number of apprentice positions. Thus, they effectively control the number of members. Professional groups such as legal and medical organizations often restrict entry into their professions by limiting the number and size of accredited professional schools.

The effect of this scarcity is wage levels that are higher than competitive levels. The excess supply of workers shows up as the number of people who try to get into the *training programs* but cannot. Because fewer people can enter the field, there is no unemployment among those who do succeed in getting in.

Now compare the two approaches to labor market control: supply restrictions and threats of strikes. They have similar results compared to the competitive market: A higher wage rate; fewer workers hired; and other workers who cannot get the job they want. Nor are the two approaches mutually exclusive: A craft union will often both restrict entry and threaten to strike.

Raising demand It is sometimes argued that unions may indirectly increase the demand for members' services, offsetting to some extent the restrictive effects of their actions. If a greater sense of security and participation causes union members to work harder and more efficiently, their marginal product curve—and therefore the demand curve for labor—will shift to the right. For professional groups, stringent quality requirements may increase the public's confidence in the quality of services, thereby increasing the demand for these services.

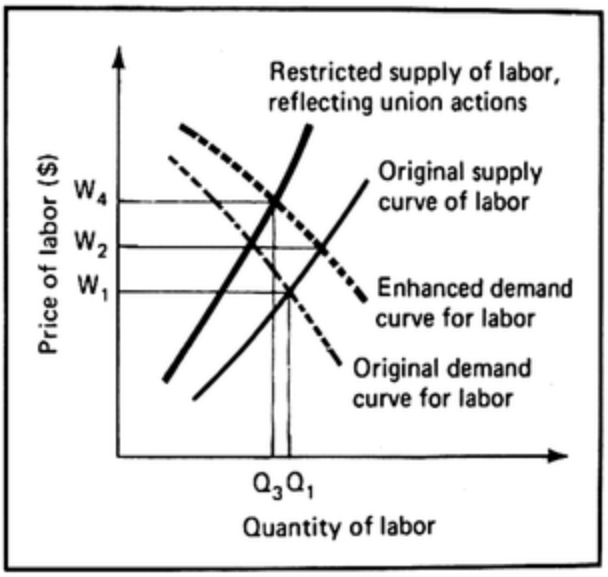


Figure 11 Union actions may raise the demand for labor

Though it restricts supply, the craft union may also take actions that increase the demand for its members' labor. The net effect of this decrease in supply and increase in demand will be higher wages. The effects on the quantity of labor hired are uncertain. Employment will rise if the demand shift dominates and fall if the supply shift dominates. In this example, the supply shift dominates, and the quantity of labor hired decreases slightly from Q_1 to Q_3.

The shifting out of the demand curve is shown by the "enhanced" demand curve in Figure 11. Both the supply reduction and the demand increase will increase wages above the competititve level. The net effect on the quantity of workers hired will depend on the relative size of the supply and demand shifts. If the supply decrease dominates, the equilibrium quantity of labor will fall below the competitive level. If the increase in demand for labor dominates, the equilibrium quantity may lie above the competitive quantity.

So far, the interferences with competitive labor market outcomes that we have discussed all come from the labor or supply side of the market. The market power, however, may be concentrated on the

buyer or employer side of the market. This is discussed in the next few sections.

Monopsony and competitive supply

Monopsony occurs on the demand side of labor markets when there is only one buyer of labor. The key difference between a monopsonist and a competitive buyer of labor is found in the slope of the supply curve that each type of firm faces. Since the monopsonist firm is the only buyer of labor in the particular market, the labor supply curve faced by the firm *is* the industry supply curve. Therefore, the monopsonist faces an *upward-sloping* supply schedule of labor. This upward slope indicates that the monopsonist can hire more workers only if it will pay a higher wage to all workers. *The marginal cost of the labor curve, therefore, lies well above the supply curve for labor.* The gap between the marginal cost of labor curve and the labor supply curve represents the amount by which the wages paid to previously hired workers increases.

The monopsonist follows the same general rule as other profit-maximizing firms: Hire the input up to the point where its marginal revenue product equals its marginal cost. The result of monopsony is that the firm pays each worker an amount *less than the marginal revenue product of the last worker.* The monopsony condition drives a wedge between the wage rate and labor's value to the employer. Not only the wage rate but also the quantity of workers used will be lower than in a competitive situation.

Pure monopsony is unusual. Even a pure *monopolist* seller may not be a *monopsonist* as an employer. Thus, an electric utility may have a monopoly on the provision of electricity to a particular area but be only one among many employers of skilled electricians in the area. On the other hand, a *competitive seller* may have *monopsony* power in hiring. For example, a furniture company may have only 3 percent of the national furniture market but still be the one big employer in a small "company town."

More commonly, monopsony power is only moderate, with two or three firms dominating the hiring in a labor market. If the oligopsony power creates extra profits, workers are likely to organize to recapture some of the surplus for themselves. The result is a labor market in which both the supply and demand sides are controlled. This situation is called a ***bilateral monopoly.***

Bilateral monopoly

When both the suppliers and buyers of labor have some degree of market power, each side hopes to use that power to reap extra benefits for itself. The result, of course, is a continuing struggle between the two sides for the largest share of the financial pie. The firm wants to pay a *low* wage rate; the union wants to obtain a *high* wage rate. The gap in between is the zone of contention.

Figure 12 illustrates the union-monopsonist battleground. The monopsonist tries to equate MRP and the marginal cost of labor, hiring Q_2 workers. It would like to pay W_2, the absolute minimum wage necessary to obtain Q_2, indicated by Point M on the supply curve for labor. The union, on the other hand, seeks the highest possible wage (W_3) for Q_2 workers, indicated by Point U on the MRP (or demand curve) for labor.

The gap between W_3 and W_2 becomes the battlefield. Each side tries to get a wage as close to its goal as possible, but will settle for nothing less than its minimum condition: W_3 for the firm, and W_2 for the union. Both sides know the range of dispute, with the union demands at the top end of the bargaining range and the company's counteroffers at the bottom. Each

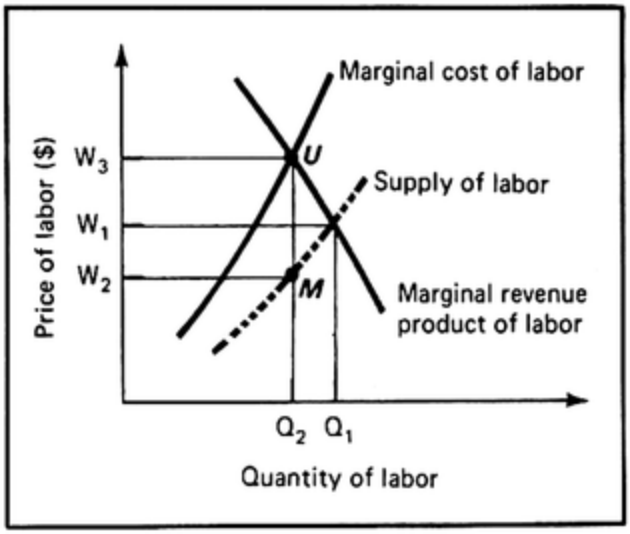


Figure 12 Bilateral monopoly between a union and a monopsonist

The monopsonist wishes to pay the low wage of W_2. The union seeks the higher wage of W_3, knowing that the monopsonist would pay it if it were forced to do so. The actual outcome depends on factors such as bargaining strength and the economic climate.

side also knows the other's strengths and weaknesses. The outcome depends on the relative bargaining power of the two sides, and so the stage is set for industrial strife, with threats, bluffs, and intense bargaining. All of the conditions that give the union its degree of leverage will now come into play. The union may threaten and actually carry out a strike.

The results depend also on the firm's ability to withstand union pressure and apply pressure of its own. The firm's power to make and carry out threats depends on such conditions as the inventories that can be sold during a strike, easy access to other workers who can replace strikers, and the firm's financial resources.

By chance, the outcome could be the same wage rate as a strictly competitive market would have given, such as W_1 in Figure 12. W_1 is close to the middle of the bargaining range set by W_2 and W_3, so that it would result from roughly equal power between union and management sides.

The effects of labor groups on workers' incomes

Many of the unions and professional groups formed during the last century have become powerful, often winning large raises. Yet, the wage increases might have occurred even without union activity.

Have unions raised wages in general significantly above what they might otherwise have been? The difficulty lies in separating out the unions' unique effects from the many other economic changes occurring at the same time.

Research indicates that unions' effects are substantial. During the 1960s, the consensus is that unions raised their members' wages by about 25 percent, compared to wages received by nonunion workers. The more recent detailed patterns shown in Tables 4 and 5 suggest that the effects have continued. Despite some variations, union wages were 19 percent above nonunion wages in 1977 (Table 4). Also, wages rose faster during 1970—1979 in unionized than in nonunionized industries (Table 5).

At the bottom of the job scale, the gains from unionization are marginal at best. These unions tend to be fairly weak because demand for their members is highly elastic, as already noted. In the middle range of the wage scale, unions appear to do best (1) when they are industry-wide, and (2) when they face firms that have monopoly power as sellers. Toward the top of the job scale, some craft unions and professional groups have raised their members' incomes by large amounts: doctors and lawyers are prime examples, along with electricians and plumbers.

On the whole, then, many labor groups have directly raised their members' wages, more so in skilled than in unskilled jobs. But these *direct* effects have also created *indirect* effects that reduce other workers' wages. Throughout this chapter,

Table 4 Comparing wages for union and nonunion workers in major sectors, 1977

Industry	Average Weekly Earnings, Union	Average Weekly Earnings, Nonunion
Mining	$296	$339
Construction	343	230
Manufacturing	243	234
Durable goods	252	250
Nondurable goods	227	214
Transportation, communications, & public utilities	293	267
Wholesale trade	255	263
Retail trade	228	175
Services	250	206
Public administration	279	258
All industries	262	221

Source: U.S. Department of Labor, *Earnings and Other Characteristics of Organized Workers, May 1977 (Washington D.C.: U.S. Government Printing Office, 1979), p. 32*

you have seen that the wage-raising effects reduce the level of hiring in the industry, so that some workers become unemployed. They then seek work elsewhere, shifting the labor-supply curves down in the other industries that they enter. The result is to reduce wages in *nonunionized* industries.

This indirect reduction of nonunion wages needs to be compared with the direct raising of union members' wages, in judging whether union activity has raised the general level of wages in the economy. The net result depends on the elasticities of supply and demand in the various industries. Shifts clearly do occur, but their relative amounts are still debated.

It is also relevant to consider the amount of work time lost by strikes, to place in perspective the impact of union activities. In most years, strikes actually occur in less than 4 percent of negotiations. Strikes have cost about 30–50 million lost workdays each year during recent decades. This, however, has been a tiny fraction of all working time, never above 1 percent and usually below ½ of 1 percent. About 4 percent of the work force is involved each year in strikes, but most

Table 5 Unionization and wage rises in selected industries, 1970–1979

Industry	Percent Unionized	Percent Wage Increase, 1970–1979
Railroads	99	130
Basic steel	98	150
Major vehicles	98	115
Cigarettes	95	124
Laundries	29	89
Knitting mills	26	83
Clothing stores	11	75
Banks	8	62

strikes last less than two weeks. Moreover, the number of major strikes (those involving 10,000 or more workers) has been declining steadily since 1970, when there were 34 in the United States. In 1978–1981, there were on average only 11 such strikes each year.

Summary

This chapter has explored the functioning of labor markets, examining influences on both the supply of and demand for labor. The key points of the chapter are summarized below.

1. Workers try to maximize the satisfaction gained from their work by balancing the benefits of work (enjoyment and pay) against the cost (dissatisfaction).
2. Even given the differing natures of jobs and individuals, working more hours will, at some point, involve decreasing satisfaction. Therefore, people usually need the incentive of higher pay to put forth more work effort.
3. There are two opposing influences on a labor supply curve: a *price effect* and an *income effect.*
4. To obtain the market supply curve of labor, individual labor supply curves are added horizontally. Since most are on the upward-sloping portion of their supply curve, the market summation of individual curves usually slopes up rather than back.
5. On the demand side of the labor market, firms will hire labor up to the point where the contribution to the value of output of the last worker hired (or marginal revenue product) equal the wage cost.
6. The intersection of the market supply and demand curves for labor will determine the equilibrium wage and quantity of labor hired in a particular market. Therefore, to understand wage differences from one labor market to the next, it is necessary to examine how supply and demand conditions in the various labor markets differ.
7. One important factor in explaining wage differences is the investment in human capital embodied in each worker. Higher amounts of training or investment in human capital are correlated with higher earnings.
8. The scarcer a particular type of labor is, in relation to demand, the higher the wage it will receive. While there is a *natural scarcity* of many talents, *artificial scarcity* may also exist.
9. There are important cases in which labor markets depart from the competitive model. Suppliers of labor can form employees' organizations such as unions to gain some monopoly power. Buyers of labor may have a monopoly on the buying side of the market, or *monopsony.*
10. A labor union's power depends on two key factors: its *control* of the labor supply, and its *leverage* or ability to inflict economic damage on the employer through such actions as strikes.
11. A monopsonist will hire labor up to the point where the marginal cost of an additional worker just equals the marginal revenue product. The result is that wages and the quantity of workers hired are forced below the competitive level.
12. In a bilateral monopoly, both suppliers and buyers of labor have some monopoly power. The result is a

struggle between the two sides for the larger share of the financial pie.

13. The actual effects of labor unions on workers' income is unclear, although some unions have made fairly significant gains for their workers.

Key concepts

Marginal utility
Price effect
Income effect
Blue-collar jobs, white-collar work
Human capital
Unions
Professional groups
Monopsony
Bilateral monopoly

Questions for review

1. Does the principle of comparative advantage apply equally to the selection of part-time jobs and lifetime jobs? Explain.
2. Two doctors graduate from medical school together and set up a joint practice. Over the years, as their real income rises, one doctor works longer and longer hours, while the other doctor takes longer and longer vacations. Using the concept of the labor supply curve, explain why this different behavior might occur.
3. Discuss the impact of each of the following on the demand and/or supply curve for labor.
   a. A firm modifies its assembly-line techniques in a way that makes labor more productive.
   b. A union manages to gain a wage increase for its members.
   c. The cost of capital increases relative to the cost of labor.
4. List three jobs that you consider to be relatively high paying and three that you consider to be relatively low paying. Try to identify some of the factors that would contribute to the wage differentials among these specific occupations.
5. What is meant by the statement that the monopsony condition will result in workers being paid "less than they are worth." Illustrate and explain.
6. Using a diagram, show the bargaining range that might exist in a bilateral monopoly.
7. Define the terms *price effect* and *income effect*. How do they interact to determine the amount of work that a person is willing to do?
8. How do the strategies of industrial unions differ from those of craft unions or professional organizations?
9. What determines the likelihood of a successful strike by an industrial union against a company? What advantages do the union and the employer enjoy against each other?
10. In which of the following types of jobs might a union provide the most gains for its members? Why?
    a. janitors
    b. dentists
    c. electricians
    d. steel workers
    e. carpenters
    f. waiters
    g. apple pickers
11. Which of the following statements are true and which are false? Explain your answers.
    a. Strikes are a major cause of lost work time in the United States's economy.

b. In the late 1970s, wages were higher in non-unionized industries than they were in unionized industries.
c. Between 1970 and 1979, the income of railroad workers, most of whom are unionized, rose faster than the income of bank clerks, most of whom are not unionized.
d. The number of major strikes that occur each year in the United States is on the rise.
e. The effect of strong unions is to raise wages throughout the economy.
f. Skilled labor is in a more advantageous position than unskilled labor when it comes to collective bargaining.

16

Capital, Investment, and Technological Change

As you read and study this chapter, you will learn:

- how firms decide how much capital investment to undertake
- what the determinants of the returns to capital are
- how the market values capital assets
- how innovations occur

To explain capital, we begin with running shoes. A shoe is a piece of capital: long-lasting and a source of improved human efficiency. Almost any shoes will do for a beginner, but a serious runner will usually invest in a more expensive, specialized pair. The higher cost will be justified by the benefits (speed, comfort, pizazz) gained in race after race.

Indeed, runners invest in more than shoes. They buy special shorts, tops, warm-up suits, key holders, complex timing watches, and so on. And as these pieces of capital wear out, they need to be replaced, just like industrial capital. Running equipment may be scant compared to golfers' club sets and motorized carts or to football equipment. Even so, runners' investments exhibit many of the essential features of capital.

This chapter explains the nature of that capital, of the returns to capital, and of investment choices that create capital of all kinds, from running shoes to factories. There are two general forms of capital: ***real*** or ***physical capital,*** such as machinery and

buildings; and ***money*** or ***portfolio capital,*** such as bonds and stock certificates. These two kinds are sharply different in some respects, but they are related, and their quantities are governed by similar kinds of investment choices that compare costs, returns, and risks. Therefore, we offer an integrated coverage of the basic issues, pausing where needed to point out the differences between real and money capital.

In the first main section, we present the nature of real capital and the basic investment choice. Next, in the second main section, we analyze the return to capital, focusing on risk and return. We also contrast interest and profit. Then, in the third main section, we show how the values of assets respond to various influences in markets. We also explain how capital markets direct industrial activity throughout the economy. We conclude by reviewing the main ways in which technological progress alters capital and its productivity, in the fourth main section.

Capital and investment decisions

What is capital?

All physical capital shares several characteristics. First, it is *productive;* it aids labor in production, so that total output is higher than it would be without capital. Second, capital is *made by people.* It is not a natural resource but a product of human effort. Third, capital *lasts over time* rather than being used up immediately. Fourth, capital involves a *present cost* to create the capital goods, but then it delivers a stream of *future values* by improving production.

Therefore, capital results from investment, which creates a set of equipment and other productive facilities. Production is diverted from present consumption goods to produce capital goods, which then raise the capacity to produce both kinds of goods in the future. **The process occurs over time, with an initial commitment of resources (the investment), followed by a period of higher productivity (the return on the investment) while the capital lasts.**

As always in economics, there is a comparison of cost with benefits. The cost is the sacrifice of present consumption as the capital is produced and installed. The benefit is the increase in the value of production that the capital brings about. For investment in capital to be efficient, the benefits must exceed the costs.

How capital aids production can be seen in several examples. At the simplest, a worker can drive nails and break stones more effectively with a hammer than with a rock or a stick. The hammer may cost a carpenter $10 to buy, but over years of use, it saves labor time and reduces the carpenter's cost by a larger amount. Similarly, a computer that requires $20 million of resources to produce may cut its owner's costs by $10 million per year for many years.

The benefits created by capital reflect the ***productivity of capital*** in use. They are called the ***return to capital.*** Returns are necessary if the capital is to be provided in the first place. If the return to capital is too low, then the benefits do not justify the cost incurred in creating it.

Actual capital and investment

The real capital of an economy is a complicated array of productive facilities, from computer chips to bulldozers, printing presses, oil rigs, and office buildings. It comes in nearly infinite varieties and can be seen almost everywhere that production occurs. Besides durable items like machinery, capital includes the inventories of

parts and finished products that businesses need to maintain their production.

On the consumer side, too, capital is extensive and familiar. Houses, furniture, appliances, even clothes, are "consumer durable goods," which is one category of capital. Besides these physical forms of real capital, the investment in human skills also creates human capital, which we discussed in the previous chapter.

About one sixth of national output is spent yearly on investment in business capital. Approximately half of the investment replaces worn-out or obsolete capital. The rest is new investment, which adds to the capital stock. Because investment decisions are crucial to future production, we must now analyze them.

The decision to invest

The decision to invest is a commitment to the future.

Whether to invest The firm has first to decide whether to invest at all. The cost is the outlay to acquire and install the capital. The benefits are the future stream of net returns (or profits) gained from selling the goods that are produced by using the capital. Future returns must be discounted for time, to reflect the fact that present dollars are more valuable now than future dollars. Panel I, of Figure 1 illustrates the effect of discounting where the *constant money stream of benefits translates into a decreasing stream of present values of benefits.* The present values are obtained by the formula

$$\text{Present value} = \frac{\text{Future value}}{(1 + r)^T}$$

where r is the interest rate used to discount future values and T is the number of years into the future. The more distant benefits are discounted more heavily, because the yearly discounting process operates over more years. A $1 million profit expected next year has a present value of $920,000 when discounted at 8 percent; the $1 million expected in the 10th year has a present value of only $463,200; after 15 years only $315,200, and so on. These

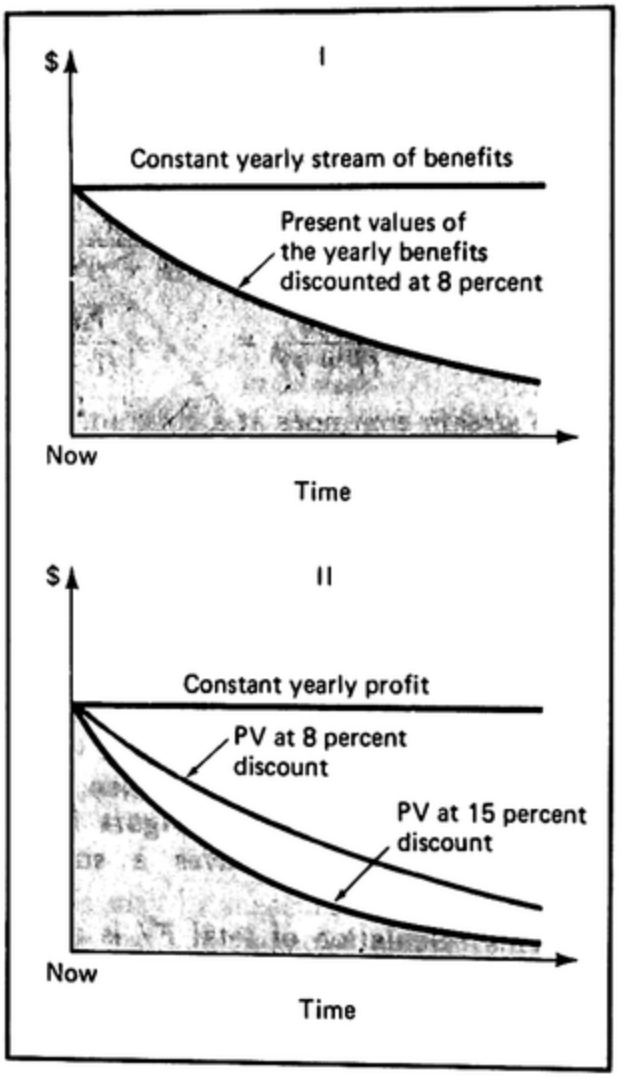


Figure 1 Discounting future payments to present values

In Panel I, discounting for time reduces the present value of a future payment below its nominal value. The effect of discounting increases as the interval increases. The sum of the present values, shown by the shaded area, is the total present value of the entire stream.

In Panel II, a higher rate of discount reduces the present value even further, leaving a smaller shaded area representing present value.

discounted present values (shown by the shaded area in Panel I) are summed to obtain the total benefits (or profits) from this amount of investment. The total PV from Period 1 forward is:

$$\text{Total present value} = \frac{\text{Profit}_1}{(1+r)} + \frac{\text{Profit}_2}{(1+r)^2} + \frac{\text{Profit}_3}{(1+r)^3} + \ldots + \frac{\text{Profit}_n}{(1+r)^n}.$$

If the stream continues at a constant level for an indefinite number of periods, then the formula becomes much simpler:

$$\text{Total present value} = \frac{\text{Yearly profit}}{\text{Discount rate}} = \frac{\text{Yearly profit}}{r}.$$

Thus, a $1 million yearly stream discounted at an 8 percent rate has a total present value of $12.5 million. A higher discount rate gives a lower present value; for example, if r is 15 percent, then PV is $6.7 million. In Panel II, of Figure 1, the higher discount rate leaves a smaller shaded area.

This calculation of total PV is called ***capitalizing the value of future benefits.*** It will be used often in this chapter.

As for the firm's choice, the expected total PV must be at least as great as the cost of the investment, or else the investment will not be worth undertaking.

How much to Invest Where investment returns are high enough to warrant some level of investment, the firm must decide *how much to invest.* The firm compares the cost and benefits of alternative amounts of investment. It arrays its investment projects from the most profitable to the least profitable. The task is to select the optimal level of investment, which is the level that will *maximize the firm's profits in the long run.* The array is based on the *rates of return* expected from the investment levels, as in Figure 2. The rate of return for each project is the interest rate *that will just equate the capitalized value of an investment's future profits with the investment's initial cost.* It is called the ***internal rate of return*** of the investment project. For example, a $10 million investment might generate $3 million yearly in profits for an unlimited number of years. By the simple equation

$$\text{Investment cost} = \frac{\text{Yearly profit}}{\text{Rate of return } (r)}$$

we have

$$\text{Investment cost} = \$10 \text{ million} = \frac{\$3 \text{ million}}{r}.$$

Transposing the $10 million and r terms, we get

$$r = \frac{\$3 \text{ million}}{\$10 \text{ million}} = 30 \text{ percent}.$$

The project, therefore, has a rate of return of 30 percent. At an interest rate of 30 percent, it will just break even. For comparison, suppose that a $13.8 million investment yields a yearly profit of $3.78 million. Its internal rate of return is ($3.73) ÷ ($13.8 million) = 27 percent. If the profit stream is for a finite period, the calculation is only a bit more complicated.

The firm ranks its possible investment projects according to the rates of return they offer, starting with the highest, as in Figure 2. What the projects accomplish often varies widely, but the investment decisions all reduce to the same basic financial terms: their dollar amounts and their rates of return. Thus, the best project requires $15 million and has a 23 percent rate of return; the second best also costs $15 million, and has a 20 percent rate of return; and so on.

Together these projects comprise an investment opportunities schedule, showing the *marginal returns on investment* (MRI). The marginal return on investment is the rate of return provided by the incremental dollar of investment; here it is ap-

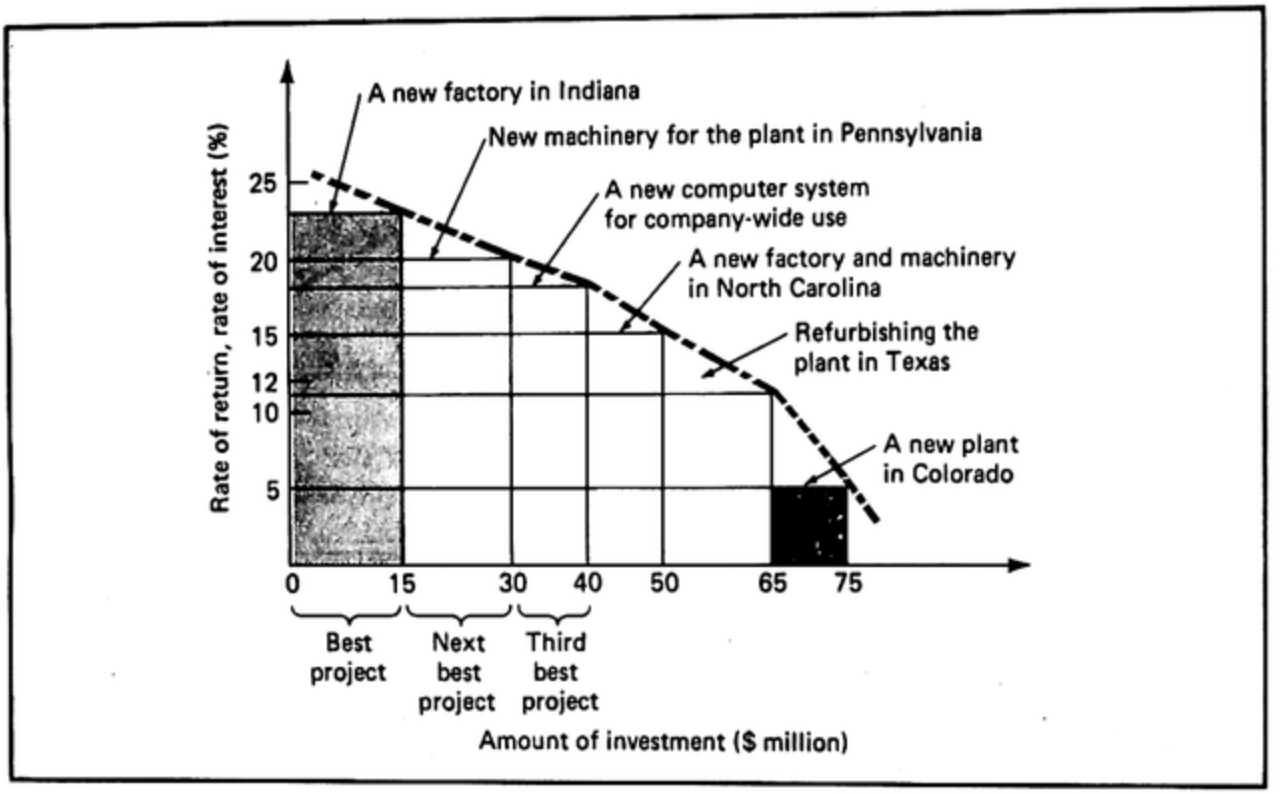


Figure 2 The productivity of capital, shown by the returns on investment projects for a typical large firm

The firm's possible investment projects are arranged in descending order of yields, starting with the best one at the left. The range of possible yields is typical of many firms' actual prospects.

proximately shown by the return to the marginal project. Thus, the firm's marginal rate of return on investment in the $50–$65 million range, where the Texas project is on the margin, is 11 percent.

For simplicity, the MRI curve is usually drawn as a smooth curve, as in Figure 3. Smooth or not, the curve shows the marginal returns available to the firm at various investment levels. Knowing this curve of possible *benefits* from investment, the firm has half of the information it needs to select the ***profit-maximizing level of investment.*** The other half is the *costs* of investment.

The cost of capital *The cost of real investment is the opportunity or market cost of the funds that are committed to the investment.* By spending its funds on capital, the firm forgoes the returns that could be obtained on the money in other ways during the life of the investment. These alternative returns may be measured by the rate of interest paid in financial markets, if financial investment is the most profitable alternative to real investment. The firm in Figure 3 could have put the funds into interest-bearing bonds rather than into real capital. The cost of capital is, therefore, the forgone rate of interest. To get the benefits of the investment projects, the firm must incur the economic cost of not earning interest on the funds. If it does not have funds, it must raise them on the market and pay the market rate of interest.

This ***cost of capital*** (CC)* in the simplest case, then, is the market interest rate. Figure 3 illustrates a 12 percent rate: All investment by the firm will cost it 12 percent a year. Because the cost of capital is

*The "cost of capital" here is the cost of the funds used to purchase capital goods, not the price levels of the capital goods themselves (such as $150,000 for a machine).

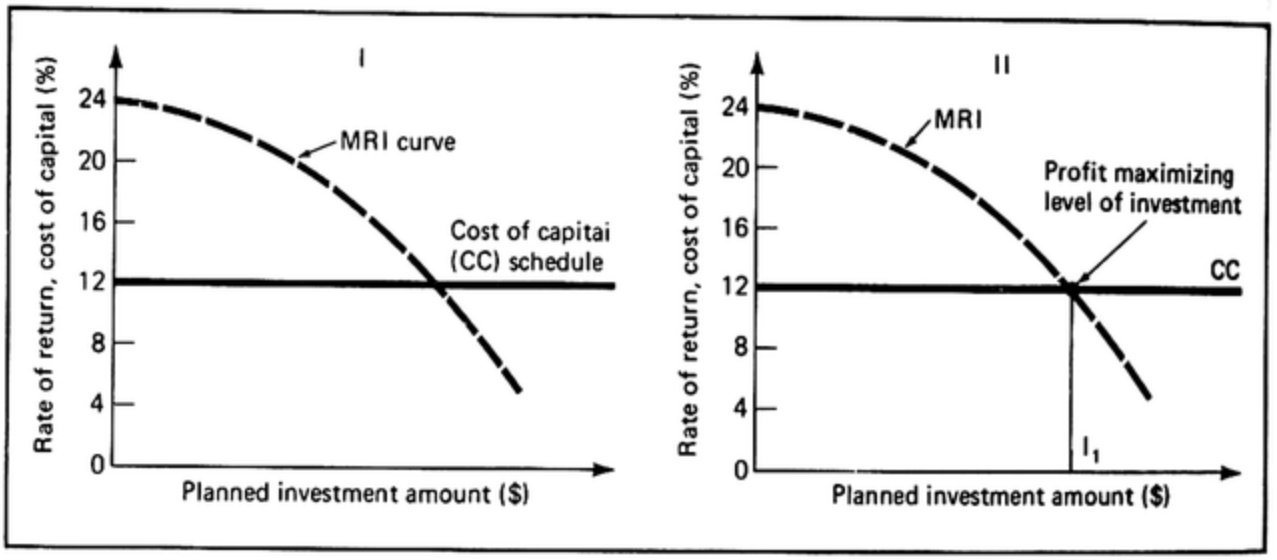


Figure 3 The returns and costs of investment

In Panel I, the marginal returns to investment curve is down-sloping, while the cost of capital is shown as constant rate of interest.

In Panel II, the profit-maximizing level of investment is at I_1, where the marginal returns to investment just equal the marginal cost of investment.

constant in this case, both the average and marginal cost of capital are equal at all investment levels.

Profit-maximizing investment A firm that has all the needed information can perform the basic benefit-cost comparison at the margin. All projects whose rates of return exceed the cost of capital will be carried out, for they add net profits. Projects whose returns are below the cost of capital will be avoided. *Therefore, investment will be set at the level where the marginal return equals the marginal cost of capital.* That is level I_1 in Panel II of Figure 3, where MRI cuts the CC curve. At that level, the internal rate of return at the margin (which equates the capitalized value of future profits with the investment's initial cost) equals the market rate of interest. The firm's profitability on the marginal investment is equated to the general cost of capital as measured by the market interest rate. The firm thereby avoids investing too little or too much.

Changes in the returns or the cost of capital will usually alter the firm's profit-maximizing level of investment. A rise in the MRI schedule will increase the level of investment whose returns exceed their cost, such as to I_2 in Figure 4. A rise might occur if the market's growth unexpectedly

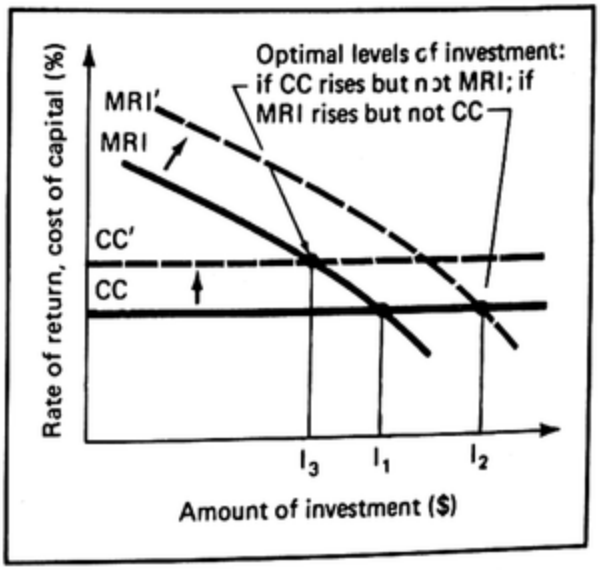


Figure 4 Effects of shifts in investment returns and costs

If the cost of capital rises, CC will shift up, causing the optimal investment level to fall to I_3. But if the returns to investment rise while the cost of capital stays unchanged, investment will rise to I_2.

increases or if the firm develops new innovations. As for CC, a rise in the market interest rate will raise CC and shift down the margin of profitable investment. In Figure 4, a rise of r to 15 percent reduces the profit-maximizing level of investment to I_3.

In general, firms' investment opportunities and the cost of capital interact in this way to set investment levels in markets throughout the economy. A deterioration of investment prospects (shifting MRI curves down) reduces investment; an improvement in prospective returns raises investment. A rise in r induces firms to invest less; a fall in r evokes increased investment.

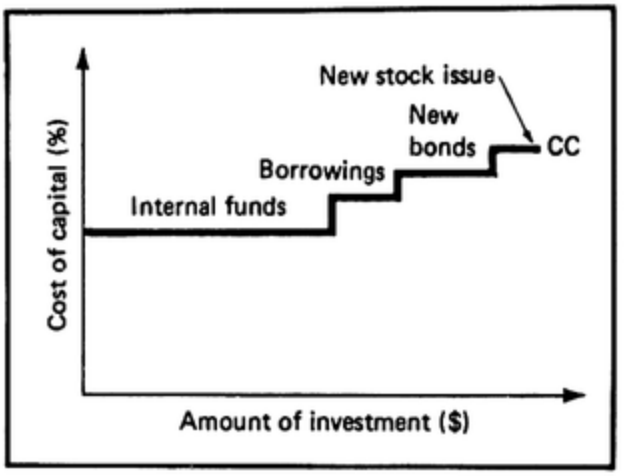


Figure 5 The firm's cost of capital curve may slope up

The firm arrays its sources of capital in order of increasing cost. Internal funds from retained earnings are cheapest, followed by borrowings and bonds, and then new stock issues.

Market demand and supply for capital

Demand *Each firm's MRI curve is its demand curve for capital.* MRI expresses the benefits of investment to the firm. Individual demand curves can be summed horizontally to obtain market demand curves for capital, providing the prospective returns to any one firm appear to it to be independent of the investment undertaken by others.

Supply The supply of investment funds to the firm is embodied in its CC curve. The simplest case is a horizontal line at the going interest rate.

But the supply of capital to the individual firm may be upward-sloping rather than horizontal. A firm has two sources of funds. One is *internal funds*, which arise when the firm reinvests its own profits. Such funds may superficially appear to be costless to the firm, but they have an opportunity cost equal to the rate of interest that could have been obtained by investing the funds outside the firm. *External funds*, obtained from lenders and investors outside the firm, have an explicit price. The cost of external funds is, therefore, the direct interest and dividend payments that must be made to bondholders and stockholders to acquire their funds.

External funds customarily have a rising marginal cost. The two kinds of external funds are *debt* and *equity*. Debt involves interest payments for loans and bonds. Interest payments are recorded as business costs and are tax deductible. Equity capital involves common stock that is paid dividends. Because dividends are paid out of profits, they are not a cost and are not tax deductible. The cost of equity capital is thus higher than the cost of debt.

In obtaining funds, the firm arrays its sources from the cheapest to the costliest, as in Figure 5. Usually it must pay lenders or investors at higher incremental rates to obtain more funds, partly because of ***risk***. Therefore, the CC schedule slopes up, as illustrated. The slope and position may differ among firms, but the same basic analysis applies to them all.

To simplify matters, the firm's CC schedule can be drawn as a smooth curve, as in Figure 6. This curve crosses the MRI schedule and determines the profit-maxi-

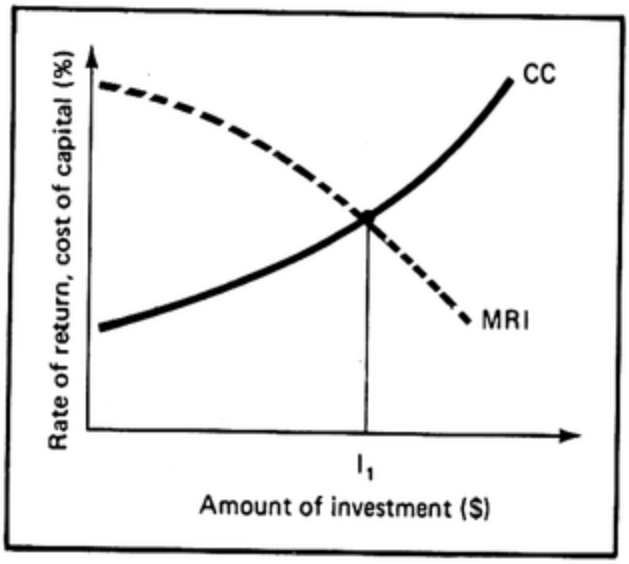


Figure 6 The profit-maximizing investment choice

When CC slopes up, the optimal investment choice still follows the same logic, with the outcome here at I_1.

mizing level of investment at I_1. The underlying logic of the choice is the same whether CC is up-sloping or horizontal.

The supply of funds to an individual *firm* varies according to how much funding it can generate internally and the terms on which it can arrange debt and equity financing on the market. For any given *industry*, the supply curve is similarly composed of an internally generated portion and a higher-cost portion representing outside financing. For a given firm or industry in sound financial health, the cost of equity or debt financing may, over a wide range, be independent of the amount of funds raised. But the supply of capital to the *economy as a whole* is obviously not independent of the amount raised, since savers cannot supply limitless amounts of new funds at the going rate of interest. The supply of funds as a whole is a complex matter, dependent on the level of national income and the elasticity of saving to changes in the rate of return. The MRI for the entire economy is also complex, since the return to investment in any one industry depends on overall prosperity, which, in turn, is linked to the rate of investment in other industries. The determination of total investment is one of the central concerns of macroeconomics, the study of the economy as a whole.

The return to capital

When we examine the return to capital more thoroughly, we find that it contains several distinct economic elements: (1) *the pure interest rate on invested capital*, which would be earned if investment had no risk; (2) a *risk premium*, which is an additional return required to reward investors for risky investment; and (3) any remainder, which is *economic profit*.

Because risk is a key concept, we discuss it next. Then we consider interest and profit separately.

Risk and return

When funds are invested, the owner bears some risk about the outcome. Since risk is normally viewed as a hazard, investors must be rewarded with extra returns to bear the unpleasantness. There is, accordingly, a ***risk-return relationship***, which is a basic feature of investment decisions. To analyze it, we must first define risk.

Risk Taking risks means accepting a probability that things will turn out badly. Life is permeated with risks, whether you are crossing a busy street or choosing a career. Some risks are voluntary, such as in skydiving. Others are unavoidable. For example, you must choose *some* occupation, and any such choice incurs risks.

For firms and investors, risk is financial: the probability of losing the value of the investment. Investment decisions involve committing funds to make profits. But the profits are not guaranteed. There are varying degrees of risk that the outcome will be worse than expected.

Risk is a matter of *probability* of loss. It is often expressed in percentage terms that reflect the odds. Thus, a project's probability of failure may be 1 percent (one in a hundred), which may be regarded as a low degree of risk. A higher percentage, such as a 10 percent or 50 percent likelihood of loss, would indicate a higher degree of risk. The familiar betting odds can be stated as probability values; "even odds" means 50 percent risk; 10 to 1 odds against a loss means a 9 percent risk.

Financial risks are also of both kinds, voluntary and involuntary. *Voluntary financial risk* is undertaken when a firm or investor selects a high-risk investment instead of a low-risk alternative. *Involuntary risk* is faced by every firm with funds to invest: *Some* choice must be made, and all alternatives involve some risk, however small. The risk may be of *commission*, as when a risky investment collapses, causing the capital to lose its value. Or the risk may be of *omission*, as when the firm mistakenly misses a chance to make a large gain.

For example, a firm with $60 million to invest may choose the "riskless" alternative of putting the funds in government bonds, whose money value will stay constant. But this may mean passing up other investments that, though they carry a slight risk of loss, offer some chance of a $30 million capital gain. Therefore, the seeming "riskless" choice (the bonds) actually bears a risk of sacrificing the $30 million in capital gain.

The simple analysis of risk focuses on the risk of loss, as a matter of probability. The future outcomes are dispersed around their most likely values, in what is called a *probability distribution*. For example, the two investments in Figure 7 may be most likely to return a $3 million profit each year if one could make only one best estimate for each one. But the possible outcomes can be much higher or lower.

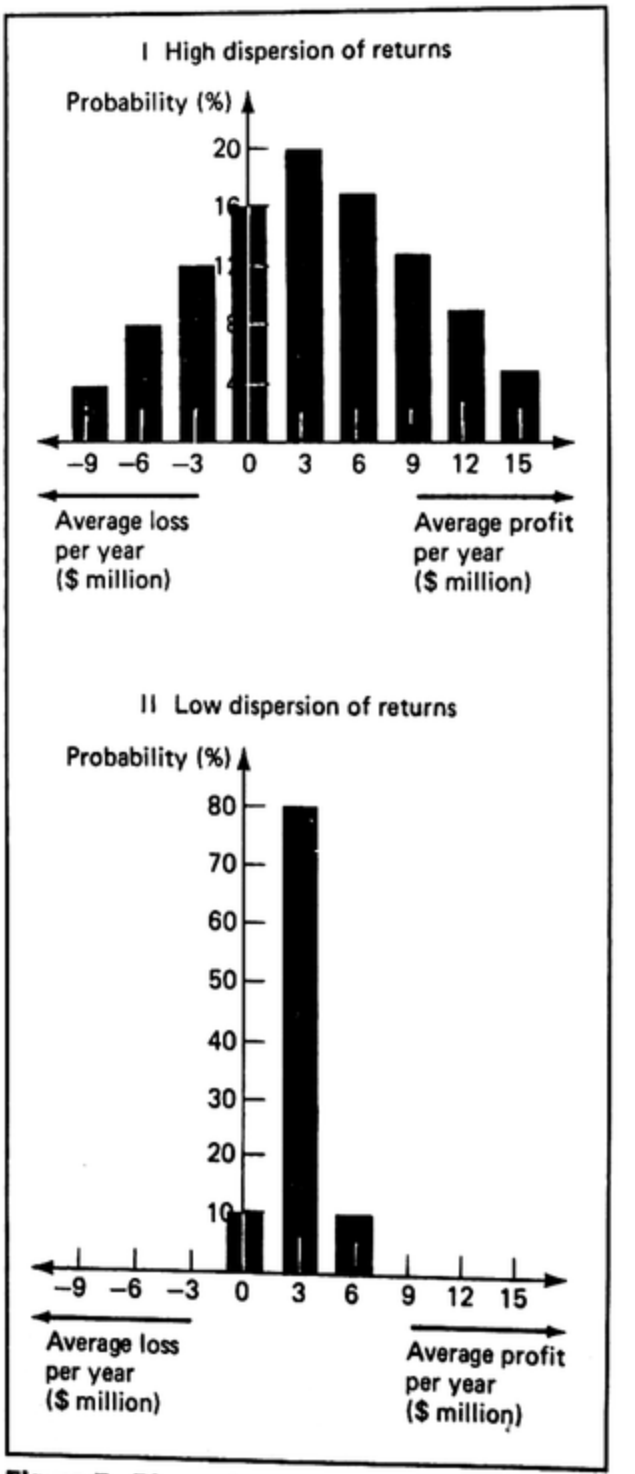


Figure 7 Dispersions of likely profits from two differing investments

The probability of each profit or loss level is the percentage likelihood that it will occur. Thus, in Panel I, a $3 million loss is expected to occur 12 percent of the time. In Panel II, a $3 million profit is 80 percent likely.

The dispersion of expected results is much wider in Panel I than in Panel II, even though the average expected value ($3 million) is identical in both.

In Panel I of Figure 7, the likelihood of an actual loss is the sum of probabilities of incurring a $3 million, $6 million, or $9 million loss—which is, respectively, 12 percent plus 8 percent plus 4 percent (the shaded bars), which equals 24 percent in total. By comparison, another project might also have a $3 million yearly profit, according to the best estimate, but with much less dispersion in its probability distribution, as in Panel II. Here there is no probability of an actual loss. Therefore, the risk exposure is much less in this project than in the first.

One definition of *risk is the share of the probability distribution that falls in the range of financial losses.* For any given level of expected profits (such as the value of $3 million in both panels of Figure 7), a wider dispersion will result in a larger probability of loss because the shaded portion that falls in the loss range will be larger. Therefore, *risk is commonly associated with the degree of variation in returns around their average expected value.*

The risk-return relationship Because risk is unpleasant, people usually will accept it only if they expect compensating rewards. For example, steeplejacks and oil-drilling workers usually get higher pay than people who have safe jobs. Similarly, in investment choices, the reward for risk bearing is a higher-than-average level of expected return.

Consequently, economists and financial analysts apply the concept of a *risk-return relationship: Investments with higher risks must offer higher average returns.* Most investors are *risk averse*—they dislike risk and must be compensated for bearing it. Only the prospect of an unusually high reward will induce them to accept danger. Exceptions can be found, of course, but they *are* exceptions.

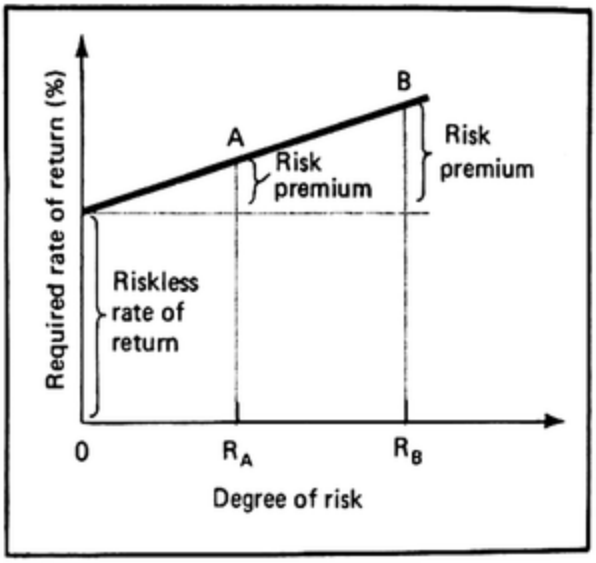


Figure 8 The risk-return relationship
As risk rises, investors require a risk premium in compensation. The riskless rate of return is expected to be earned, on average, by all investments. The total rate of return includes both the riskless rate of return and the risk premium.

Accordingly, there is usually a positive relationship between risk and expected returns. In Figure 8, risk is on the horizontal axis, from zero to high degrees of risk. The required rate of return is on the vertical axis. As risk increases, so does the rate of return that the investor will require.

At zero risk, the required rate of return will be 10 percent, as shown. That rate of return would be available without appreciable risk by investors on short-term U.S. government securities. At higher risk levels, higher rates of return are required. The difference between these rates of return and the ***riskless rate of return*** is the *risk premium:* the reward for bearing risk. In Figure 8, a risk level of R_A involves a specific risk premium, as shown; the higher risk level of R_B requires a risk premium twice as large. These premiums reflect the general preferences of investors between risk and returns, as expressed by their actions in financial markets.

Alternative outcomes along the risk-re-

turn relationship are equivalent to one another for the average investor. At each point, the risk premium offsets the disadvantage of the risk. Therefore, the riskless return is as attractive to the average investor as Point *A* with R_A risk or Point *B* with R_B risk.

Of course, *individual* investors and firms may have differing risk-return preferences. Some dislike risk intensely; others are only mildly averse to it. Extremely risk averse investors will tend to choose lower-risk investments, toward the left, unless they are compensated by very high returns. Those more willing to accept risk will make choices more toward the right, without demanding a big risk premium.

Regardless of these variations, the basic risk-return relationship is a central part of investment decision making, both for firms and for investors.

Risk-return patterns can be observed in actual financial markets. Risk itself is difficult to measure, because it deals with future probabilities. Those are inherently unknowable with precision, as are all future conditions. One measure of risk is to be found in the ratings of bond quality that are made by the two major investment analysis companies, Moody's Investors Service and Standard & Poor's Corporation (S&P). The ratings reflect the estimated probability of default on the bonds, which would reduce or eliminate their value. In making the ratings, the two firms assess many features of the companies and governments that issue bonds: their stability, prospects, profitability, and other conditions. They then assign the bonds a rating ranging from "high quality" to "speculative," as follows:

	High Quality		Investment Grade		Substandard		Speculative
Moody's	Aaa	Aa	A	Baa	Ba	B	Caa to D
S&P	AAA	AA	A	BBB	BB	B	CCC to D

The ratings reflect judgments, not mechanical formulas. They are widely accepted as useful practical indexes of risk.

They also directly affect the interest rates that must be paid. A B-grade bond will have to offer a rate of interest higher than the rate on an Aaa bond, precisely because the B bond carries more risk of loss. U. S. government long-term bonds have the lowest yields because they have the lowest risk.

Diversifying One way to reduce risk is to choose a single low-risk investment, such as a U.S. government or Aaa bond. Another way is to *diversify*, by choosing several investments that, though they individually have risk, will tend as a group to offer stability. Known colloquially as "not putting all your eggs in one basket," diversifying deserves explanation.

The fundamental principle is the *law of large numbers*: The larger the sample drawn from a random distribution of possible outcomes, the more likely it is that the sample's average value will equal the total distribution's average value. In short, there is safety in numbers. *The yield on a diversified selection of investments will tend to be close to the average yield of all assets in the market.* It will be higher than the yield on a single risk-free investment.

With the risk-return relationship clarified, we now turn to the analysis of interest and profit.

Interest

Interest is a contractual payment to compensate lenders for the use of borrowed

capital. The borrower obtains the funds either as *loans* from specific lenders (such as banks) or by selling *bonds* to the market. The borrower pledges to pay a fixed amount of interest per year to the lender (or bondholder). Calculated as a percent of the loan (or bond's) initial value, the money payment is the rate of interest.

For example, a firm may borrow $30 million either by taking a loan from a bank or by selling $30 million worth of bonds. If the rate of interest is 10 percent, the firm will have to pay $3 million in interest each year besides having to pay off the $30 million borrowed. Any failure to pay that $3 million will place the firm in legal default and perhaps ultimately in bankruptcy.

The interest rate that a firm pays reflects the going rates of interest in financial markets as a whole. Those rates, in turn, reflect the productivity of real capital, the pattern of consumer and business saving, the government deficit, the state of the business cycle, and a host of other macroeconomic variables.

Variations among the rates that different firms must contract to pay should be interpreted as risk adjustments. The risk-free rate of return is pure interest. Any excess beyond this is compensation for the likelihood of default.

Profit

The total return to the owners of a firm—sole proprietors, partners, or corporate stockholders—is all called profit. As noted, profit has three components: (1) a pure or riskless rate of return, (2) a risk premium, and (3) economic profit (the remainder).

The first two components of profit (riskless return and risk premium) are similar to the components of interest paid on loaned capital. They *must* be paid if investors are to supply equity capital. Thus, a firm might have $178 million in net income after taxes, but the riskless and risk components account for $82 million of it. The remaining element—economic profit—would be $96 million. Economic profit is large when the firm is successful, but low or negative when the firm is in difficulty.

Economic profit can be subdivided into two further components of its own. First, it is *a return on entrepreneurship and innovation* by the firm's managers. Creative policies develop new products and put the firm into new lines. For such excellence, part of economic profit is an appropriate reward.

Second, the firm may hold *monopoly power,* and *part of the profit may be a payment for that advantage.* The size of that element will depend on the firm's degree of monopoly power. For example, Firms A and B have identical profit rates of 25 percent on equity capital. Yet, Firm A is highly risky and has no monopoly power, while Firm B has little risk, strong entrepreneurship, and much monopoly power. Actual cases offer endless variations.

Total profits are available for the shareholders' benefit. Part of profits is usually paid out to shareholders as *dividends;* in recent years, the dividend pay-out ratio has averaged between 30 and 50 percent. The rest of profits are *retained earnings,* which are reinvested in the firm. They enlarge the firm's capacity to produce, increasing its earning power. Therefore, retained earnings usually increase the value of the stock, by making it more likely that its price will rise and provide a capital gain when it is sold. In contrast, dividends provide direct money benefits to shareholders.

Total corporate profits have a long-term growth trend, but fluctuate with the business cycle. Their pattern in recent years is shown in Figure 9. The share of profits paid out in dividends declined dur-

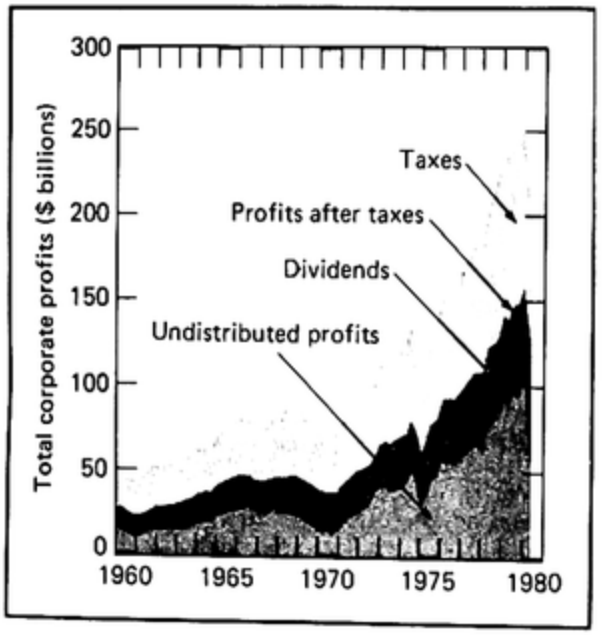


Figure 9 The trend of total corporate profits

Profits have grown as the total value of production has risen. The share of profits paid out in dividends has declined, while the share of undistributed profits (retained earnings) has risen.

Source: Federal Reserve Board, *Historical Chart Book.*

ing the 1970s from about 50 percent to about 30 percent.

The value of assets

There are well-developed markets for all kinds of physical and portfolio capital. Each item of capital commands a value or price in the market. That value can change, as economic forces alter the equilibrium price. Both physical and portfolio capital can undergo price changes. No matter how solid a factory may be or a gilt-edged bond certificate may look, its market value can rise or fall, perhaps sharply.

We now show how economic processes determine those ***asset values.*** The processes are driven by ***expectations*** and apply both to real and to paper assets. They enable the stock market to function as a system of control that limits the actions of all corporations in the economy.

But first we show how sharply asset values do, in fact, fluctuate. We focus on stocks and bonds, but pay some attention also to real assets.

Fluctuations in asset values

Stocks are the most widely known fluctuating asset. Each corporation issues stock, which is then bought and sold by private investors in stock markets. The price of each company's stock at each moment reflects the supply and demand for that stock. They, in turn, reflect (1) the company's specific conditions (its profits, future growth, etc.), and (2) general shifts in the whole stock market (when the whole market is rising or falling, all stocks are affected).

Over 1 million corporations' stocks are traded, but the largest several thousand firms are the main focus of the major stock exchanges, as noted in the box on page 344. Several stock market averages have been developed to measure stock price swings: Figure 10 shows two of them. The "Dow Jones Industrial Average" covers just 30 leading firms, while the NYSE Composite Index is broader, covering all 1,900 stocks traded on the New York Stock Exchange. These indexes can be followed virtually minute by minute each weekday, as they respond to market changes.

The fluctuations are substantial. Between January 1973 and September 1974, the NYSE average fell from 63 to 36, a drop of 43 percent. It then rose by even more; between late 1974 and late 1979, it more than doubled (the 30 firms in the Dow Jones average did not rise as much). Then the NYSE index fell again by 22 percent between November 1980 and March 1982.

The Main Parts of the Stock Market

Decisions to sell or buy stock are made by millions of investors. Most transactions are carried out through stockbrokers, acting in one of these major stock markets:

Stock Exchange	Number of Companies Included	Main Types of Firms Included	Volume of Shares Traded on an Average Day
New York Stock Exchange (New York)	1,900	Largest and next largest firms, banks, utilities, etc.	50,000,000
American Stock Exchange (New York)	800	Large and middle-sized firms in all sectors	5,000,000
"Over the counter" market	3,500	Small firms; new and risky firms	27,000,000

The price of each transaction is publicly noted. Any broker can obtain it almost instantly. The stock prices and trading volume for about the leading 3,000 companies are published the next day in the *Wall Street Journal;* other papers focus on the larger firms.

About 20 million people own stocks, but only 200,000 or so hold $1 million worth or more each. Most of the remaining investors hold less than $5,000 worth of stock apiece. These are the "small" investors. "Institutional" investors (large private holders, insurance companies, pension funds, bank trust departments, etc.) now make over 70 percent of all stock transactions. These professional investors' operations also vary in size. Perhaps the largest 300 of them make over one half of all stock-market transactions.

Many *individual* stock prices moved much more sharply, however. Four examples are given in the two panels of Figure 11: Sears and Polaroid dropped by 70 percent or more, while Time Inc. dropped by 60 percent and then quadrupled, and Holiday Inns rode an even steeper roller coaster. Many more such cases can be found in standard stock market reference sources.

A swing in the stock market can change asset values by hundreds of billions of dollars. Thus, the total value of all corporate stocks fell by several hundred billion dollars during January 1973–September 1974, then rose by about $600 billion during 1974–1979, and dropped $200 billion during 1980–1982. These changes have a very large impact on the wealth of over 20 million stock owners.

Bonds also fluctuate in value. Because bonds provide fixed interest payments, they have long been regarded as having stable asset values. Yet, sharp changes in

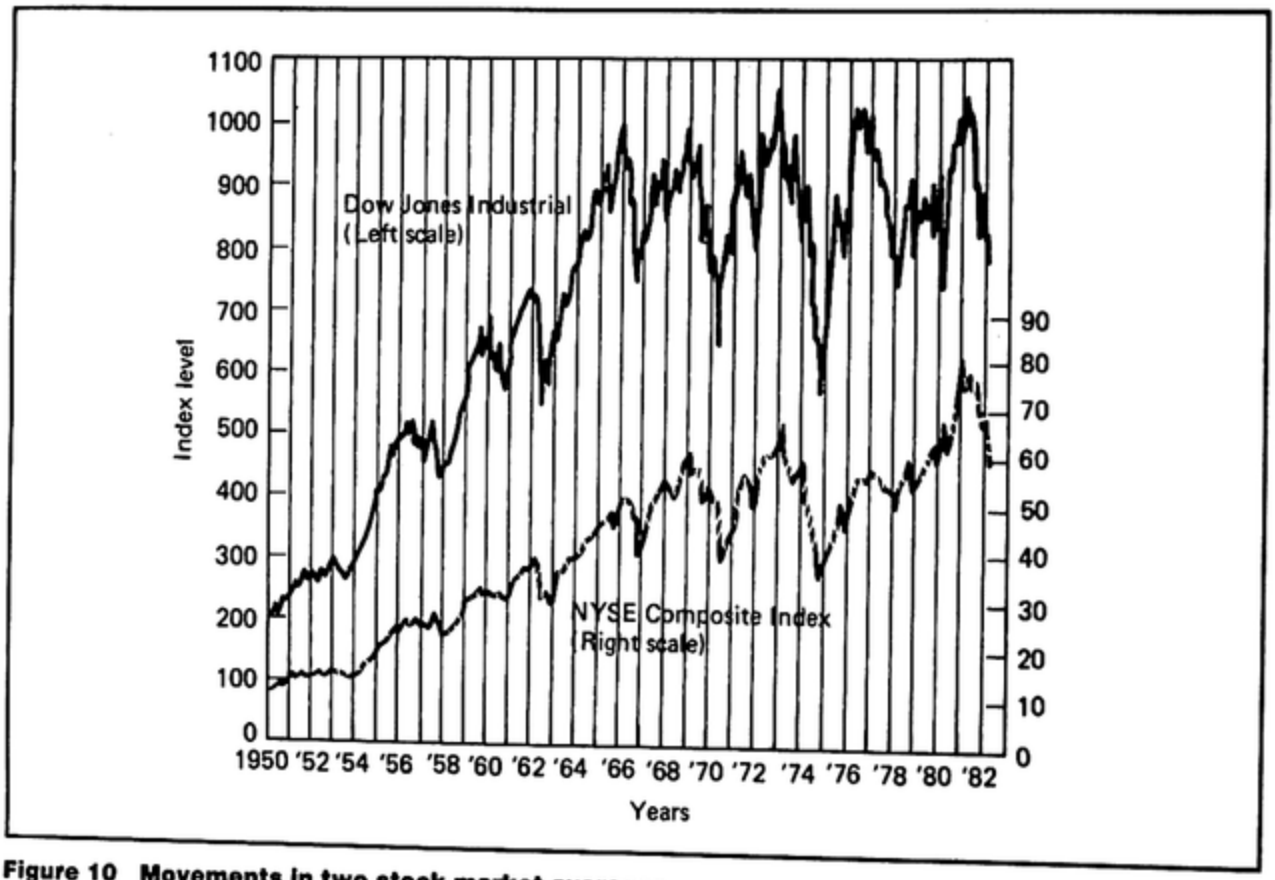


Figure 10 Movements in two stock market averages

The Dow Jones Industrial Average covers 30 leading industrial firms, while the NYSE Composite Index includes all stocks traded on the New York Stock Exchange. Their movements are largely parallel, but there was some divergence during 1979–1982. Both indexes indicate the sharp fluctuations that occur.

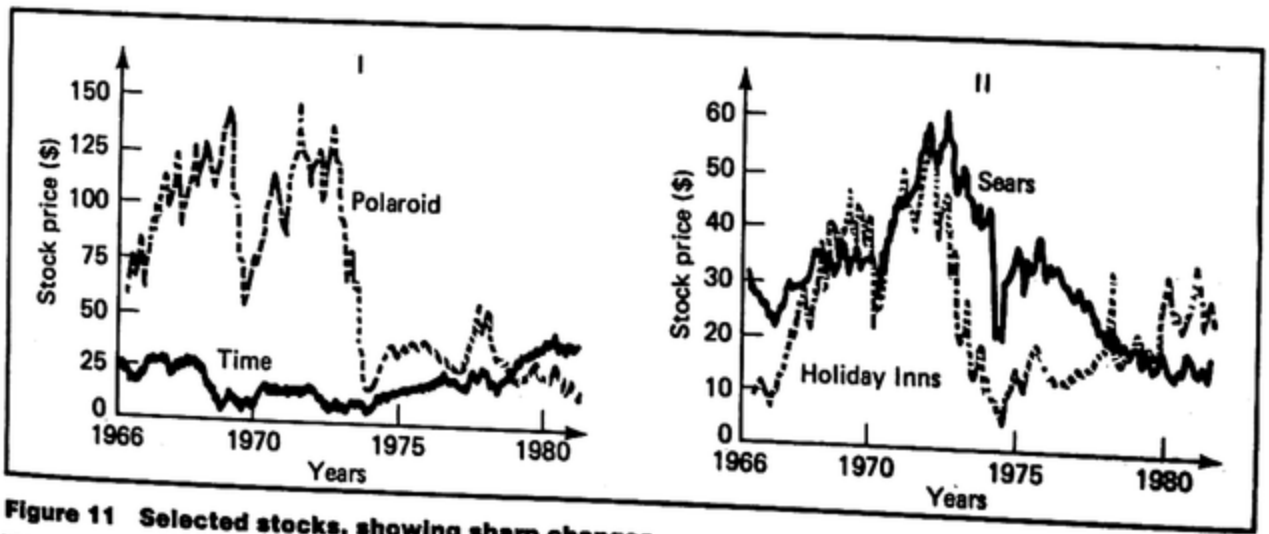


Figure 11 Selected stocks, showing sharp changes

These four stocks have shown sharp price changes since 1965. Values have changed by as much as a multiple of 5, often in short periods. The changes reflect conditions internal to the firms, as well as various industry-wide factors *and* changes in stock markets as a whole.

bond values have occurred since 1965 (for reasons that will be explained shortly). The prices of many bonds fell by over 40 percent during 1977–1982.

Real assets (for example, machinery, office buildings, houses, oil tankers) also fluctuate in value. Office buildings provide a good illustration. The demand for office space can increase sharply. Because the supply is fixed in the short run, such booms raise the rents and prices of existing buildings. New construction then occurs, sending rents and building prices down again. The physical characteristics of existing buildings do not change, but their prices do.

The forces causing such large changes in values are strong and pervasive. Superficially, one can cite shifts in demand and supply as the causes. But what causes those shifts? The answer is a single crucial phenomenon: expectations.

Expectations govern asset values

Expectations about future returns and interest rates are the main determinants of asset values. They are forward looking and reflect human judgment about the probabilities of future events. The market values of assets differ from accounting values, which are based on past expenditures.

Expectations affect prices through the process of investment choice. Each asset (real or portfolio) offers a flow of expected benefits. These benefits are capitalized, by discounting them for time. If risk is high, a risk premium is also appropriate, and the future values are discounted not only for time but also for risk.

The logic of asset valuation is universal: It applies both to physical capital (which is used to produce real goods that may be sold at a profit) and to portfolio capital (which is held to receive interest or dividend payments and possible capital gains).

In the simplest case, the future benefits are a constant stream continuing for an indefinite period. Then the simple formula for capitalizing the stream to a present value can be applied. The discounting rate is the going rate of interest for assets with the given degree of risk. To illustrate, consider again the asset providing a $3 million yearly profit flow, at a time when the market rate of interest on assets bearing that degree of risk is 10 percent. The present value (PV) is

$$\text{PV} = \frac{\text{Yearly profit}}{\text{Interest rate}} = \frac{\$3\,\text{million}}{10\,\text{percent}} = \$30\,\text{million}.$$

Because 10 percent is the opportunity cost of the capital, this investment must pay a rate of return at least as high. Therefore, the market rate of interest sets the "required rate of return" on the asset. And that, in turn, sets PV, as you can see.

The present value is, in fact, what the asset is now worth. The value will change if either element in the ratio changes. A rise in expected profit levels will raise the asset's value proportionally. For example, if expected profits jump to $4 million yearly for any reason, the asset's value will rise to $40 million (which is $4 million ÷ 10 percent).

Interest rate changes will also affect the asset's value. Suppose profits stay at $3 million, but the interest rate doubles to 20 percent. The asset's value will sink to $15 million (which is $3 million ÷ 20 percent). If the interest rate then falls to 5 percent, the asset's value will *rise* to $60 million (which is $3 million ÷ 5 percent).

These asset values are not capricious. They are the equilibrium values that demand and supply in capital markets will actually yield. At its creation, an asset's value is governed by the cost of providing it. Once established, its market value is dependent on expected returns and interest rates. Fluctuations in asset values can be very sharp, not only in stocks but also in

bonds, productive capital, and other assets (gold, paintings, cattle, houses, etc.). Sufficiently unfavorable changes can reduce an asset's value to zero, even though its physical and legal characteristics are unchanged.

Bonds and stocks

Now consider bonds and stocks specifically in more detail. *Bonds* offer two kinds of future yields: a stream of *interest payments* plus a possible capital gain (or loss) in selling the bond.

For simplicity, we consider only very long term bonds, which will be redeemed in 25 or more years. That interval is so long that the current prices are virtually unaffected by the future redemption value. Such bonds are free to fluctuate widely in price.

They do fluctuate. The recent rises in interest rates have driven bond prices sharply down. In a typical case, a $10,000, 40-year bond was issued in 1965, paying $550 per year (5.5 percent interest at the start). The rise in the interest rate to 11.5 percent has caused the bond's price to fall to $4,800, with an effective interest yield of the same 11.5 percent (which is approximated by $550 ÷ $4,800). Thousands of other bonds' prices have moved similarly. Of course, if the bond is bought now at $4,800 and held while interest rates fall to their 1965 levels, then it could be sold for $10,000 and provide a capital gain of 108 percent (which is [$10,000 − $4,800] ÷ $4,800).

Stocks are slightly more complicated to analyze because their dividends are subject to change. The most common expectation is that a firm's dividends will grow as the firm's sales and profits grow. The expected future dividends are estimated by beginning with the recent year's dividend payments and then factoring in a growth factor, which can be labeled g.

To illustrate, suppose that AT&T's recent dividend is $6.00 per year, and that this dividend is expected to grow at 5 percent per year. The expected AT&T dividend in any future year (designated as Year T) is then the original dividend plus a factor for 5 percent growth during the interval from now to Year T. For example, the expected dividend next year (when T is 1) is $6.00 $(1 + g)$ = $6.00 (1.05) = $6.30. The cumulative growth in the dividend level by the fifth year will be

$$\begin{aligned}\text{Expected dividend in Year 5} &= \$6.00\,(1+g)^T \\ &= \$6.00\,(1+g)^5 \\ &= \$6.00\,(1.05)^5 \\ &= \$6.00\,(1.2763) \\ &= \$7.66.\end{aligned}$$

Suppose that the growth factor g is constant: The firm's future growth in profitability and dividends is expected to continue steadily. Then g can be included in the simple present-value formula as follows:*

$$\text{Present value of the stock} = \frac{\text{Current dividend}}{\text{Interest rate} - \text{Growth factor}} = \frac{D}{r-g}$$

The greater the expected rate of dividend growth (g), *the greater will be the stock's value* (for any given levels of D and r). That is logical because a stock is simply the right to the stream of dividends; more dividends means that the stock is more valuable.

In the example, if the interest rate is 15 percent, then the present market value of AT&T stock would be

$$PV = \frac{\$6.00}{.15 - .05} = \frac{\$6.00}{.10} = \$60.$$

*The proof for this role for the g factor can be found in standard texts on managerial finance.

Sometimes the current *dividend yield* on a stock (dividends as a percent of the stock's current price) is below the market rate of interest. In the above illustration, the current $6.00 dividend pays only 10 percent on the current $60 value of AT&T stock, while the interest rate is 15 percent. But that simple comparison neglects the growth factor. Because dividends are expected to rise, the current yield is below the expected total rate of return. The rise in dividends will cause the share's price itself to rise by 5 percent yearly over time, so that it can be sold for a capital gain. The stockholder's total return, therefore, includes *both a dividend yield based on* r *and a capital gains yield based on* g.

The choice process equalizes returns

The process of investor choice will bring returns on stocks into line with those on bonds. This is known as the ***equalization of returns.*** *The process tends to equalize rates of return throughout the capital markets (portfolio and physical).* If equilibrium is reached, all investments of equivalent risk offer equivalent returns. And among investments of differing degrees of risk, all rates of return will differ just enough to provide compensating risk premiums.

Three levels of knowledge

The equalizing process does not, of course, operate with steel-trap finality. Markets are often out of equilibrium. Thus, there are many opportunities to make unusually high (or low) returns. But to make high returns one must have superior knowledge on three levels:

1. *Real conditions:* within firms (new products, planned growth, etc.) and financial markets.
2. *Financial effects:* expectations about how the firms' financial performance (profits, dividends, rates of return, etc.) will be affected by the real conditions.
3. *Timing:* how far *other* investors have already acted. It is crucial to act *before* others do; otherwise the potential gains are all discounted away in advance. Timing is crucial.

Therefore, it is not only *what* you know but *when* you know it that matters. You must not only know the firm and its financial prospects. You must also outguess the rest of the market by acting before other investors have reaped the possible gains.

For example, suppose that the Eastman Kodak Company develops a new type of camera and film, which will probably raise its growth rate of profits and dividends from 5 to 10 percent per year. An investor needs to know that real and financial information, but the third category can be even more important. Specialists working for large investors will learn of Kodak's plans almost as soon as they are made. By buying Kodak stock immediately, they can drive up the price many months before the actual innovation occurs and becomes well known.

Other investors, especially small investors, will tend to act too late. Indeed, the Wall Street adage is "Sell on good news" because the stock's price has probably been driven up *above* its new equilibrium level by the time any good news about a company reaches the newspapers.

Stock markets as a control system

By performing well, a firm's managers create growth in profits, which usually leads to rising stock prices. Investors, acting on expectations, bid up stock prices when firms' prospects improve or bid them down when prospects worsen. Their assessments of the future cause stock prices to change *now.*

The stock price is, in fact, a crucial index of the firm's expected performance, for three reasons. First, it reflects the firm's whole performance as a profit maximizer, as judged by skilled financial investors. Second, it looks ahead to discount future prospects into an immediate present value; there is little or no delay in applying the appraisal to present money values. Third, the stock price is of great practical concern to the firm's managers. They wish to keep investors satisfied by performing well, so that the stock's price rises and will provide capital gains to the firm's shareholders. If managers' performance levels suddenly drop, the lowered expectations will cause the stock's price to fall and thereby cut the value of the owners' investment. That stirs stockholder discontent and may ultimately jeopardize the managers' jobs. In the extreme, if stock prices fall too far relative to the value of physical assets, the firm will be bought up by outsiders and either reorganized or merged into another firm.

For these reasons, the stock market limits managers' performance. It pressures them to eliminate inefficiency, forcing costs down to the minimum possible levels on the average cost curves. It keeps them from resting on their laurels, for the market discounts results ahead of time, based on expectations about future achievements.

To the extent that it is rapid, informed, and powerful, the stock market is a system of control. Nearly all significant firms in the United States are under this baleful eye. Even if the firm issues no more new shares, the trading of existing shares maintains the pressure.

Stock markets also supply funds to businesses for investment. That role is important, but it has been shrinking since 1960, as firms have relied increasingly on internal funds and bonds rather than new stock issues. But even if stock markets provided no new funds at all, they would still exercise their control function over corporate managers throughout the economy.

Capital and technological change

The use of physical capital is closely related to technology. The current technology is embodied in the forms of capital itself. As technology changes, new forms of capital are produced and installed, replacing the old. The new possibilities for raising productivity are put into practice by investment decisions. The resulting changes in the productivity of capital have caused much of the economic progress in modern economies in the last two centuries, and in the future, they may be the only escape from increasing resource scarcities.

Trends of capital and productivity

The growth of total national output occurs partly because of simple growth in the labor and capital inputs, but it largely depends on *the progress of knowledge and skills that make both capital and labor more productive.*

Total output can grow more rapidly than the growth in inputs. To illustrate, if inputs grow at 1 percent per year while output grows at 3 percent, then ***total factor productivity*** (TFP) rises at 2 percent per year; the inputs have become 2 percent more productive each year.

In fact, TFP has probably risen at just under 2 percent annually during the last century. Though 2 percent yearly may not seem to be a high rate of gain, it cumulates over decades into very large increases. For example, 2 percent growth raises $1,000 in 100 years to $7,236. And 3 percent would raise the final total to $19,184! Even a small rise in technological progress can eventually provide a large yield.

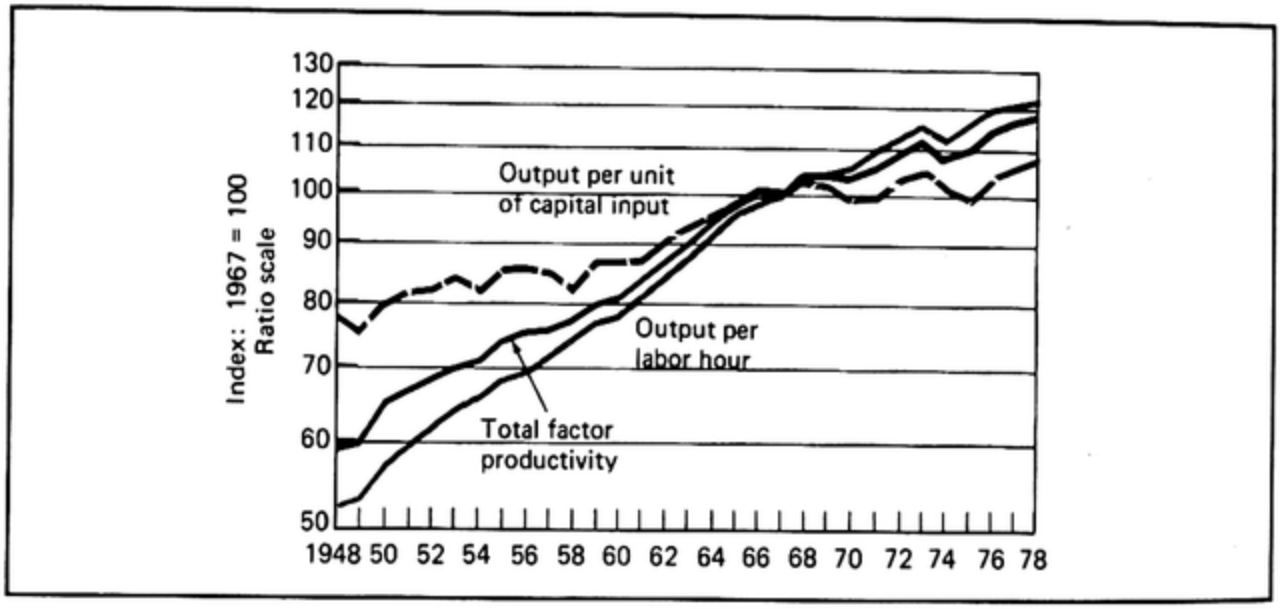


Figure 12 Productivity in the private domestic business economy 1948–1978
Sources: U.S. Department of Labor, U.S. Department of Commerce; Conference Board.

Economists therefore study productivity trends closely to discover what causes the gains and how they could be raised further. Trends for the decades since 1948 are summarized in Figure 12 and Table 1. The amount of output per unit of labor and capital rose, as Figure 12 shows. Output per hour of labor more than doubled, from 50 in 1948 to 120 in 1976. Output per unit of capital rose more slowly, from 78 to 107. The disparity partly reflects the fact that capital per worker was rising; by using an increasing amount of machinery, workers became more productive.

Table 1 ***Growth rates in total factor productivity in the U.S. economy, by sectors, 1948–1976***

Sectors	Average Annual Rates of Growth for the Entire Period (percent)				
	Output	Capital	Labor	Total Factors (a weighted average of capital and labor growth)*	Total Factor Productivity (output growth minus total factor growth)
Farming	0.9	1.1	−3.9	−2.1	3.0
Mining	1.7	0.7	−0.3	0.1	1.7
Construction	2.6	5.0	1.4	1.6	1.0
Manufacturing	3.3	3.0	0.6	1.2	2.1
Transportation	2.1	0.7	−0.5	−0.3	2.4
Communications	7.1	5.8	2.0	2.8	4.2
Electricity and gas	5.8	4.3	0.9	2.7	3.0
Distribution	3.8	3.4	1.3	1.6	2.1
Financial	3.6	5.5	2.1	3.5	0.1
Services	3.7	3.8	1.7	3.8	1.6

*The weights are based on the relative shares of the factors in total costs.
Source: John M. Kendrick and Elliot S. Grossman, *Productivity in the United States: Trends and Cycles* (Baltimore: The Johns Hopkins University Press, 1980), pp. 36–46. Copyright © 1980 by The Johns Hopkins University Press.

Table 2 The main sources of rising productivity in the U.S. economy, 1948–1976

	Contributions to the Growth of Real Gross National Product (expressed as an addition to the yearly growth rate)
Growth rate of real gross national product (corrected for inflation)	3.9%
Increase in the volume of labor and capital inputs (growth rate)	1.0%
Increase in the productivity of factors (growth rate):	2.9%
Advances in knowledge	1.4%
Formal R&D activity	0.85%
Other sources of advances	0.55%
Changes in the quality of labor	0.6%
Other changes (in the composition of industry, economies of scale, etc.)	0.9%

Source: Kendrick and Grossman, *Productivity in the United States*, p. 16.

But that was not the only influence. Both capital and labor were improving in quality. Capital was embodying better technology, and labor was becoming better educated and more skilled.

The resulting trends differed strongly among sectors of the economy. Look especially at the TFP column of Table 1. The most rapid rises were in communications, electricity and gas, and farming; TFP growth was slowest in finance and construction. Within the manufacturing sector, more detailed measures show that there was rapid TFP growth in textiles, electrical machinery, and chemicals, but slow TFP growth in such older industries as metals, machinery, and furniture.

In general, the causes of productivity growth lie in the improving quality of both labor and capital. That is indicated in Table 2 for 1948–1976. Of the total 3.9 percent yearly growth in total output during 1948–1976, only 1.0 percent was caused by the expansion of capital and labor. The remaining 2.9 percent increase in TFP was due to other causes, mainly advances in knowledge.

The trends of technology and productivity growth are the crucial sources of economic progress. We now consider their main forms and components.

Forms and components of technological change

The progress of technology is a series of individual innovations, whose development is divided into several economic stages. Once you understand them, you can analyze both the causes and value of individual innovations and the entire flow of technological change.

First, recognize that progress comes from economic effort. To develop new technology requires inventive skills and new investments, which are commonly designated as research and development (R&D).

Though R&D is mostly done by private firms, over half of all U.S. R&D expenditures are paid for by the federal government. Much of the government-sponsored R&D is for military and space programs, although some also goes for civilian products.

These R&D resources create progress in the ***invention-innovation-imitation sequence.***

Invention, Innovation, and Imitation Technical change can be divided into three phases:

> *Invention is the creation of the new idea.* The act is intellectual: the perception of a new condition, of a new connection between old conditions, or of a whole new area for action. Inventions, large or small, involve new ideas that can be refined for practical use. *Invention* is usually a lonely activity requiring intensive mental exploration. Eccentric thinkers are often best at this, although large-scale "team research" may be necessary for some large and complex inventions.
>
> *Innovation brings the idea to practical use.* The innovator builds production facilities and brings the new product or process to the market. This often (although not always) displaces products or processes in general use. *Innovation is a business act.* The entrepreneurship involved in the financing, arranging of complex engineering details, and taking of risks goes beyond the management of old processes. Though many innovations are small and safe, some require extraordinary skills.
>
> *Imitation then follows as the innovation is copied by others.* Economists call this the *diffusion* of the innovation across the market. Imitating is usually easier and safer, though less creative, than invention and innovation. The imitator merely copies, often when the innovation has become safe and routine.

Process and product Innovations Technical changes divide into two main kinds: process and product innovations. *Process* innovations alter how given products are made (examples: a new way to use a drill press, or to organize a factory, or to smelt a metal). *Product* innovations create a new good for sale (examples: a digital watch, a new kind of toothpaste, a new type of automobile, or a larger model in a line of electric motors).

The two kinds of innovation are distinct in concept, though they are often mixed together in actual cases. They call on different resources. Their incentives and effects often differ sharply. Each can vary from trivial differences to radically new approaches.

Decisions to Innovate

Since innovation requires investment in R&D and new capital, the decision to innovate resembles the decision to invest.

For *product innovations*, the incentive to bring a new product to market arises from the expected return from marketing the product. If the firm predicts that a new product will yield profits higher than the returns available from alternative uses of its funds, the product will be developed and marketed.

A *process innovation* that lowers the costs of production in an industry will be adopted by new firms that enter the industry. But existing firms may not immediately adopt the new technology. The old methods may be *technologically* obsolete, but they are not *economically* obsolete unless the firms can no longer cover the variable costs when using the old technology.

Some existing firms may be on the point of replacing worn-out machinery or equipment. They would replace their present methods with capital embodying the innovative technique. Yet, other firms would find it profitable to continue to use their existing production facilities until their variable costs are above the total cost of the new technique.

As the new, lower-cost technology spreads, the marginal costs of production and, therefore, the profit-maximizing price will fall. When the market price falls below the average variable cost of the firms using the old technology, those firms can no longer cover their variable costs, and the old technology becomes *economically obsolete.* This occurs when the average total cost curve of the new technology falls below the average variable cost curve of the old technology. At this point, firms with the old technology will make larger losses by continuing to operate than by shutting down. Therefore, all firms must eventually either adopt the new technology or go out of business.

Sources of change: Induced and autonomous ***Induced inventions*** occur from the hope of making money. Without that cash stimulus, they would happen later or not at all. Much commercial R&D activity fits this type. Teams of scientists in company laboratories, working under budgeted plans, seek inventions that will pay off for their companies: no payoff, no inventive effort.

In contrast, ***autonomous innovations*** arise spontaneously from the ongoing growth of knowledge and technology. Discoveries in one area often make an advance in another area inevitable, which, in turn, permits or causes progress in still other fields. For example, the automobile became possible only after the discovery of oil and the development of the internal combustion engine and rubber tires.

The patent system

The distinction between autonomous and induced changes can be crucial to economists in assessing public policies to promote technological change. The U.S. *patent system* is a long-established institution whose purpose is to promote inventive activity. It grants a 17-year exclusive monopoly right on an invention to the first person to file for a patent on it. The patent monopoly offers the inventor a high reward for being first.

About 70,000 patents are issued yearly, and many industries (including drugs and electronics) are strongly influenced by the race for patents and by the profits from the resulting patent monopolies. Most other countries also have patent systems, with similar terms and effects.

A patent system is economically efficient if its *benefits* (from speeding up inventions) exceed its *costs* (from monopoly restrictions on the patented goods). But the system imposes net social costs if all inventions are *autonomous;* inventions would occur even without the incentives provided by patents, but the social costs of monopoly are imposed. Only if inventions are mainly *induced* might the social benefits of the patent exceed its costs.

Therefore, the patent system's value is highly controversial. The economic appraisal depends on (1) what share the induced inventions are among all inventions; (2) whether the induced inventions are important or trivial; and (3) how much they are speeded up by money rewards. Only if the three answers are (1) large, (2) important, and (3) substantial does a patent system have a clear economic justification.

Summary

Capital is one of the most interesting factors in economic production. Some of its dimensions you will want to understand and remember are:

1. Real capital is produced and used to increase the efficiency of production over a period of years. Portfolio capital is paper securities (bonds and stocks) held for their expected money returns.

2. Actual capital is highly diverse, both in company assets and in consumer assets.
3. In decisions to invest, expected future returns are discounted to present values. Prospective rates of return are compared with the cost of capital in selecting profit-maximizing levels of investment.
4. The returns to capital contain a riskless component, a risk premium, and economic profit from other causes.
5. Rates of return usually are systematically related to risk.
6. Asset values fluctuate to reflect expected returns and interest rates.
7. Capital market decisions tend to equalize the rates of return on all assets.
8. Stock markets supervise and reward corporate performance.
9. Technological change is embodied in capital investment.

Key concepts

Real (physical) and money (portfolio) capital
Productivity of capital
Return to capital
Capitalizing the value of future benefits
Internal rate of return
Profit-maximizing level of investment
Cost of capital
Risk
Risk premium
Risk-return relationship
Riskless rate of return
Asset values
Expectations
Equalization of returns
Total factor productivity
Invention-innovation-imitation sequence
Induced and autonomous innovations

Questions for review

1. Which of the following can be classified as capital? Explain your choices.
   a. farm tractor
   b. coal
   c. football stadium
   d. 18-wheeler truck
2. Explain how a firm would derive a schedule showing the benefits of various investment projects. What is this schedule called?
3. a. How is the cost of capital investment projects determined? Is capital investment financed by internal funds costless?
   b. How would the firm use cost-benefit information to determine which investments would be most profitable?
4. Explain how each of the following would affect a firm's rate of investment.
   a. A technological breakthrough that will substantially lower investment costs requires the purchase of new equipment.
   b. The market rate of interest falls.
5. a. Would an economist view the total return on capital investment as profit?
   b. What factors determine how much profit a firm receives from an investment?

17

General Equilibrium

As you read and study this chapter, you will learn:

- how general equilibrium unifies the parts of the economic system
- the specific conditions of an efficient general equilibrium, and the meaning of the "Invisible Hand"
- how changes ripple out through the economy
- the five kinds of gaps and limits on the efficient equilibrium

Since the 1870s, the unifying concept of microeconomics has been *general equilibrium.* It reduces the great variety of entire economic systems to the simplicity of a few clear, logical relationships. Having mastered those ideas separately in the previous chapters, you can now learn them as an integrated whole. Seeing the system as a whole is one of the economist's special skills.

But there is more. Not only may a general equilibrium exist. It may also generate an efficient allocation of resources, if the markets in the economic system are competitive. That property was stated briefly in Chapter 9; now it is time to develop it more fully. Efficient allocation has been demonstrated and refined by a century of neoclassical theorists, from Leon Walras and Alfred Marshall on. Yet that efficiency does not guarantee paradise, for it often has gaps and limits.

To understand efficient allocation and its limits, you need first to grasp the interrelatedness of markets in the economy,

which the first section of this chapter shows. The second section presents the conditions of efficient allocation and discusses how the system adjusts to changes.

The last section reviews the limits on the competitive outcome and explains how social cost may diverge from private cost. Then the possible failures of the competitive process are shown in turn.

The general process toward equilibrium

Thus far, the book has offered partial-equilibrium analysis, which deals with decisions made in separate markets. In ***general equilibrium,*** the analysis widens to embrace the whole economy. Because the process of allocation operates throughout the economic system, the analysis has to be equally inclusive. All of the elements of general equilibrium have already been developed in earlier chapters. Now we combine them to derive the general patterns.

"General" means interrelated ***In the analysis, as in the economy, every part is ultimately related to every other part.*** Some markets are more closely interrelated than others, of course. Close relationships occur for many substitutes and complements. For example, coal and oil, beef and pork, automobiles and gasoline, tools and lumber, all interact closely. Changes in one such market can strongly affect the other market's outcomes. Yet, many other parts of the economy are only faintly related; New York real estate and San Francisco labor markets, for example, or beef and sulfuric acid. Yet, ultimately, they can all interact to some degree, for they all draw on the common pool of inputs and go to the same broad variety of consumers.

Any change anywhere will cause ***ripple effects*** to spread through the economy, like waves across a pool. Changes are transmitted among markets and sectors, binding the whole system together. If many changes occur at the same time, the ripples may cross and interact, so that the whole set of repercussions in the economy is mixed and seems unclear. Yet, each ripple has its own logic and can be analyzed as a distinct sequence. The economist is skilled in following such effects clearly, with a sure touch for their direction and degree of impact. We will see in the second main section that even distant repercussions can be traced. For now, the basic lesson is that ripple effects are the process by which the economy absorbs changes and adjusts to them throughout its parts.

Adjusting toward equilibrium The adjustments do not just occur randomly; they pull outputs and prices systematically toward the equilibrium patterns. The process and the interrelationships were first defined rigorously by Leon Walras in the 1870s (see the adjacent box). At each point, there are basic conditions of demand and supply that define the conditions for allocation throughout every part of the economy. The economy will adjust production toward those conditions rather than oscillate aimlessly.

MARKET CHOICES PROVIDE THAT PULL TOWARD EQUILIBRIUM If a market price is high enough to permit the firm to gain excess profits, for example, then new firms may enter the market, increase supply, and thereby drive the price down toward average cost. Each unit only maximizes its own interest. But the whole set of resulting actions causes prices and quantities in the market to adjust. The markets are cleared in each time period, so that there are neither physical shortages nor leftover surpluses. Since all goods are sold at the going price, production flows smoothly to consumers. The equilibrium conditions can continue, period after period.

The whole self-correcting process can be vigorous and thorough. If even a small gap remains, either the sellers or the buy-

General Equilibrium and Input-Output Analysis: Walras and Leontief

Although the classical economists dealt with the economic system as a whole, it was neoclassical economists who first envisioned the economy as a system of interrelated parts, thus providing the basis for a detailed analysis of general equilibrium.

The leading neoclassicist was Leon Walras (1834–1910), whose *Elements of Pure Economics* presented a comprehensive set of formulas representing all parts of the economy. By reducing the economy to a complete mathematical system, Walras was able to show precisely the fundamental processes and conditions lying beneath the great variety of the economy. And by stressing the interrelationships among markets, Walras clarified general equilibrium as the set of conditions toward which market choices adjust. Despite later refinements, Walras's analysis is still the core of modern microeconomics.

In the 1930s, Wassily Leontief (born 1906) developed the practical embodiment of Walrasian analysis, which he called "input-output analysis." Leontief, who left Russia in the 1920s, created a research organization at Harvard to process the large volumes of data that input-output tables reflect.

In later decades, he refined and enlarged the analysis, and by the 1960s, the government-issued tables presented hundreds of sectors. Leontief's work in conceiving this approach and implementing it won him a Nobel Prize.



LEON WALRAS

WASSILY LEONTIEF



Table 1 ***Conditions of efficient allocation***

1. *Markets.* All markets are cleared, with no shortages or surpluses. All resources offered for sale are used.
2. *Households*
   a. In choosing among consumer goods 1 through n, each household reaches these conditions:

$$\frac{\text{Marginal utility}_1}{\text{Price}_1} = \frac{\text{Marginal utility}_2}{\text{Price}} = \cdots = \frac{\text{Marginal utility}_n}{\text{Price}_n}$$

   for goods that it chooses to consume.
   b. In choosing how hard to work, people reach the level where the marginal utility of income just balances the marginal disutility of work.
3. *Firms*
   a. The firm's output is set at the level where

   Price = Marginal cost = Average cost (at its minimum).

   b. Input choices are made—among inputs 1 through n—so that:

$$\frac{\text{Marginal product}_1}{\text{Price}_1} = \frac{\text{Marginal product}_2}{\text{Price}_2} = \cdots = \frac{\text{Marginal product}}{\text{Price}_n}$$

   and each input is used until its marginal value product equals its price.

ers—or both groups—will take action in their own interest, pulling the outcome into an equilibrium that is part of the entire economy's general equilibrium.

When it works well, the process is therefore spontaneous and automatic. No centralized knowledge or coordination to discover the correct values and require firms and consumers to meet targets is required to make it happen. People on Nebraska farms interact independently with others in New York offices and Seattle neighborhoods, for example, and their choices tend to harmonize in a general equilibrium.

Conditions and processes of general equilibrium

In general equilibrium, resources are allocated in definite patterns throughout the economy. ***If the markets are competitive, that allocation will reach a special set of conditions that economists call efficient.*** They include the price-cost conditions that Chapter 9 showed for individual markets. They also include the input pricing conditions presented in Chapter 14. Now we assemble them all to show how they obtain across the whole general equilibrium system.

These conditions of ***efficient allocation*** are the analytical core of microeconomics. Though they are stated precisely here, they need not always hold exactly in every part of a real economy. They define the conditions that the economy is adjusting toward, whether they are reached precisely or just approximately.

The conditions

A competitive setting is necessary to obtain an efficient allocation of resources. To be competitive, markets must have many sellers, operating independently of one another and preventing any one firm from dominating. New competitors can enter quickly and easily.

In this setting, the decisive microeconomic results are *marginal.* They are reached by households maximizing their utility and by firms maximizing their profits. As they choose and adjust in individual

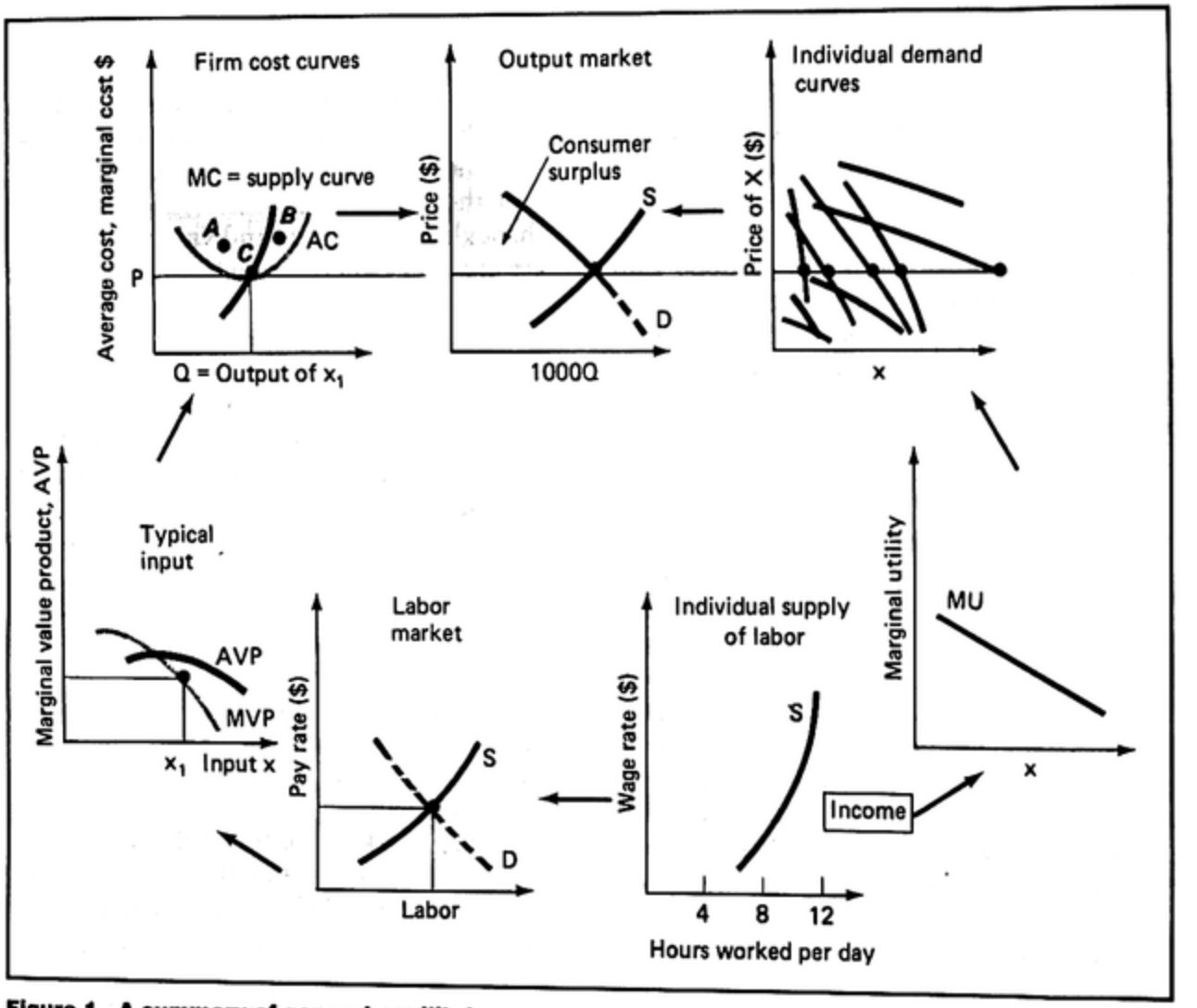


Figure 1 A summary of general equilibrium

As before, the inputs and outputs flow clockwise (recall Figure 1 in Chapter 1), while the money to pay for them flows counterclockwise. But now we see how each set of decisions involves the underlying technology (suppliers) and preferences (consumers).

Start with the output market at the top, where demand and supply come together. Demand is the sum of individual demand curves (to the right), which reflect each consumer's marginal utility (below) and income level.

Meanwhile, supply is the sum of firms' marginal cost curves (to the left), which, in turn, reflect the technology embodied in the firms' marginal and average cost curves (below). Costs also reflect the prices determined by demand and supply in input markets, such as for labor (bottom center). Finally, those labor prices govern the household incomes that flow back into the demand side on the right.

The system is therefore complete, and all parts are related to the rest. Since all of the decisions are made at *the margin*, marginal conditions are crucial.

markets, these decision units determine prices, costs, and output and input levels throughout the system. The whole outcome is governed by *marginal* choices, which tend to align values and costs everywhere in the economy.

Table 1 and Figure 1 summarize the main conditions. The conditions all fit together, and in doing so, adjust mutually toward a general equilibrium. Changes in one market will change the values in other markets, and further waves of change will then ripple through the system. The whole process of adjustment to these changes will move the economy toward these ***marginal conditions.***

The conditions are familiar to you, of course, since they are merely the partial conditions presented in earlier chapters. But this whole is more than the sum of

those parts. ***The parts of the economy interact and adjust, and increased output in one part usually requires reductions in other parts.*** Those are the special features of the general equilibrium system. In making all the changes and interactions that continue throughout the economic system, not only does one household or firm move toward equilibrium, all of them do.

Efficient allocation in competitive markets If markets are competitive, then the resulting allocation of all those resources, among all their uses throughout all of the markets, will usually be drawn toward efficient patterns. Firms in all of the competitive markets act to reach the same long-run condition: ***Price equals marginal cost at the lowest level of average cost.*** That condition was explained in Chapter 9, where we discussed partial equilibrium in a single market, but it holds generally, across all markets. *Marginal cost is true cost*, opportunity cost, the real measure of sacrifice. *Price is value*, showing what consumers will pay for the marginal unit of the good out of their own limited funds, in free choices between this and other goods.

Therefore, if price equals marginal cost then all resources are used efficiently throughout. Sacrifice and value to consumers are everywhere brought into line at the margin. Any distortion of such an efficient pattern—for example, with fewer computers than the optimum but more X-ray machines, as in Figure 2—would be inferior. In the computer market, price would exceed marginal cost: Computers might sell at a price of $10,000 but have a marginal cost of only $7,000. Since the computers would cost $3,000 less than their value at the margin, producing fewer of them would sacrifice that extra value. X-ray machines might now cost $14,000 each to make, but sell for only $11,000, as before. Producing the marginal X-ray machines would be wasteful, for their cost would exceed their value.

Both markets would be distorted from efficient outcomes, and the total value of GNP would fall below its potential level. The greater the distortion, the more GNP

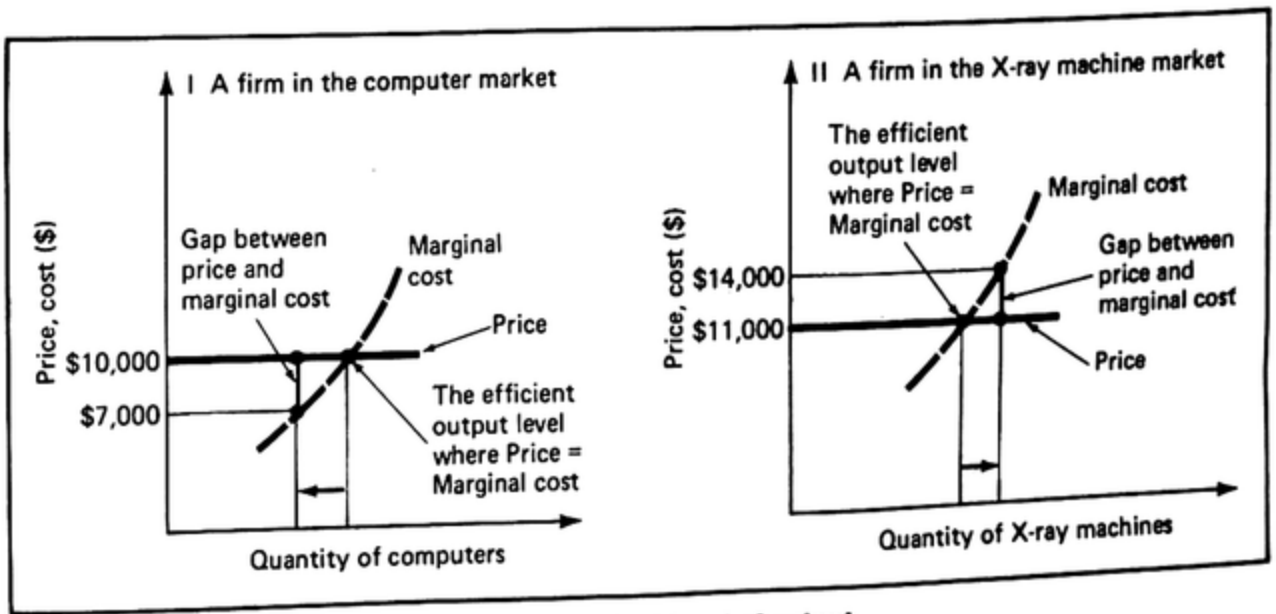


Figure 2 Marginal cost and price set the efficient level of output

The distortion arises from having too few computers and too many X-ray machines. In this situation, computers cost only $7,000 to make, at the margin, but they are worth $10,000 each, as shown by the price. X-ray machines are worth only $10,000, as before, but their marginal cost is $14,000. By moving to the levels where price equals marginal cost, there would be a net gain in welfare.

would be reduced. A simple switch of resources and production levels back to the efficient marginal conditions would again raise GNP to its maximum.

Efficient allocation is a powerful result, with large economic benefits. It maximizes national output for any given amount of inputs. Because resources are scarce, they need to be used sparingly. Consider work, for example, an input that involves personal sacrifice. When millions of people rise at 5:30 a.m. for the dawn shift at the factory, or when they endure the eighth daily hour of heavy lifting, nonstop typing, or tomato picking, they are making real sacrifices. Other inputs are also costly. For example, ores are dug, transported, and smelted, using complex capital equipment. All of these operations entail sacrifices—opportunity costs—which are embodied in the costs of the goods themselves. To waste such resources causes needless pain and effort, often of great magnitudes, measured in billions of dollars. By minimizing society's whole sacrifice in production, efficient allocation makes a substantial contribution.

Behind the simple price equals marginal cost equation lie all of the other equimarginal conditions: in input choices, consumers' ratios of marginal utility to price, and so on. They, too, would be violated by any distortion in output levels. Therefore, efficiency is a pervasive array of conditions, touches all decisions in the economy, and interacts to bind them together. Such a set of conditions sounds impossibly theoretical, but it has real meaning even in the imprecise workaday world. These ideal conditions help economists to define any large distortions that occur and to show what changes would improve economic efficiency.

The invisible hand The process works in a self-activating way, as Adam Smith noted over 200 years ago:

> As every individual, therefore, endeavours as much as he can both to employ his capital in the support of domestic industry, and so to direct that industry that its produce may be of greatest value: every individual necessarily labours to render the annual revenue of the society as great as he can. He generally, indeed, neither intends to promote the public interest, nor knows how much he is promoting it. . . . He intends only his own gain, and he is in this, as in many other cases, led by an invisible hand to promote an end which was no part of his intention.

The phrase ***Invisible Hand*** has become overused, and some economists go too far in worshiping "The Hand" and its efficient work.

Yet the process does have power. Millions of people, each acting selfishly—as consumers, workers, managers, and resource owners—may pull resources toward their best allocation, throughout the economy. A harmonious social result may arise from all of these purely private motivations. No single authority arranges it, but everybody unwittingly helps make it happen. Billions of decisions are made individually every day, and yet the whole set of markets may keep adjusting toward efficient allocation.

The process is resilient, not brittle or unstable. Disturbances are absorbed, as the process adjusts toward the next efficient equilibrium. Even where competition is not complete or other gaps arise, the basic integrity of the process remains. It also induces workers, managers, and other participants in the economy to apply great effort and care in their activities. As Adam Smith suggested, a competitive economy can be a powerful engine of economic affluence. The process is not universal; as we will see, it has gaps and limits. But where the Invisible Hand does reach, it is a strong organizing force.

Ripple effects

Now consider more closely how changes spread from market to market. The process

can be seen clearly by using an *input-output table* of the economy. A recent such table for the U.S. economy is given in condensed form in Table 2. Over 200 million people and 14 million enterprises interact in many thousands of markets. Here, they are summarized by a selection of 15 of the total 85 sectors, arrayed both at the side and at the top of the table.

As you read each row stretching across the table, the numbers show how much sales volume flowed *from* each industry to each of the other industries that are listed across the top edge of the table. *Reading each column vertically,* you can find where each industry's inputs came from, among the industries listed on the left-hand edge. Try tracing such flows for petroleum refining, following the shading of rows and columns. Most of its *output* went to chemicals ($1,800 billion), transportation ($1,999 billion), and wholesale and retail trade ($1,375 billion), as expected. Its *inputs* came mostly from crude petroleum and natural gas ($11,556 billion). Try tracing the flows for lumber products, iron and steel, and others. Once you have grasped the basic logic, the table is simple to use.

The input-output table is suggestive in three ways: (1) It helps you to visualize the many directions in which each industry relates to others, both in getting inputs and in selling its outputs. (2) It also helps you to see how any change may affect three, five, ten, and even more additional rounds of change. For example, suppose that the boll weevil cuts cotton production in half. Less cotton will flow into the textile indus-

Table 2 ***Input-output table of the U.S. economy 1967: selected sectors (amounts in $ million)***

Total Sales Going from These Sectors to These Sectors →		Oil and Gas	Raw Chemical	Lumber Products	Paperboard Containers	Chemicals and Products	Petroleum Refining	Iron and Steel	Other Metals
Industry Code Numbers	Industry Code Numbers ↓	8	10	20	25	27	31	37	38
8	Crude petroleum and natural gas	374	—	—	—	49	11,556	—	—
10	Chemical mining	—	61	—	—	608	—	26	3
20	Lumber products (except containers)	—	1	3,492	13	55	2	74	76
25	Paperboard containers and boxes	—	1	24	109	116	75	18	16
27	Chemicals and products	164	27	160	141	4,407	623	392	184
31	Petroleum refining	33	3	106	30	1,800	1,831	116	48
37	Iron and steel	120	18	43	32	184	8	6,017	340
38	Other metals	—	2	3	20	322	41	859	6,723
49	General industrial machinery	86	—	11	2	70	65	344	89
59	Motor vehicles and equipment	—	—	5	—	16	3	121	46
65	Transportation and warehousing	146	15	429	227	611	1,389	1,420	539
68	Electric, gas and water utilities	172	44	124	45	699	462	822	427
69	Wholesale and retail trade	175	16	441	151	543	303	934	726
70	Finance and insurance	93	9	79	17	100	250	182	116
71	Real estate and rental	2,429	40	109	96	543	630	80	153
	TOTAL	15,031	1,027	12,905	6,031	23,182	26,975	31,723	20,870

Source: *Survey of Current Business,* April 1979.

tries. From those industries, in turn, less output will flow to still others such as wholesaling and retailing. Meanwhile, the cotton industry will *buy* fewer inputs—machinery, gasoline, fertilizer, and so on—from other industries. Those industries, in turn, will buy fewer of their inputs from still other industries. So both the rows and the columns help you to see how changes ripple through the economy.

Moreover, the system does not change aimlessly. The repercussions follow definite patterns, which economic analysis can often clarify. Therefore, (3) the table also helps you to visualize that the responses to changes will usually pull the economy toward the efficient patterns. To see this clearly, consider some big and little practical examples of ripple effects. Always note the *net* changes—the shifts that *did* occur compared to what would otherwise have occurred.

Oil prices Let the price of oil rise sharply, as it did during 1973–1974 and 1979–1980. Then costs and, therefore, prices are likely to rise in all sectors of the economy that use oil heavily, such as gasoline and oil fuels, petrochemicals, plastics, electricity, and asphalt. If demand has any elasticity, consumers will respond to higher prices by buying less of the goods. How sharply output will fall depends on demand elasticities. Sales revenue may rise or fall, again depending on the elasticities. The rise in oil prices will also cause a fall in demand for complementary goods, such as automobiles. Thus, the industries sup-

								← FINAL DEMAND →					
General Industrial Machinery	Motor Vehicles and Equipment	Transportation and Warehousing	Utilities	Wholesale and Retail Trade	Finance and Insurance	Real Estate and Rentals	Intermediate Outputs, Total	Personal Consumption Expenditures	Federal Government Purchases	State and Local Government Purchases	Total Final Demand	Total Production (Intermediate plus Final)	
49	59	65	68	69	70	71							
—	—	26	2,521	—	—	165	14,692	—	—	—	339	15,031	8
—	—	1	—	—	—	6	838	2	—	31	189	1,027	10
22	43	2	1	92	—	29	12,118	259	30	4	787	12,905	20
14	47	26	3	571	—	4	5,841	73	34	21	191	6,031	25
13	61	46	56	215	1	301	18,797	504	1,752	113	4,385	23,182	27
30	73	1,999	275	1,375	92	720	14,105	10,194	1,078	292	12,870	26,975	31
845	103	261	41	37	—	62	30,395	4	290	2	1,328	31,723	37
308	673	66	11	42	—	37	19,752	15	8	—	1,119	20,870	38
514	297	135	37	69	—	22	4,844	—	303	13	2,956	7,800	49
41	12,157	71	3	99	—	116	15,464	15,822	1,002	731	28,276	43,740	59
71	716	5,228	563	1,235	115	1,096	32,172	11,396	3,324	985	20,653	52,823	65
52	195	342	6,888	2,415	392	436	21,370	13,935	344	1,599	15,952	37,321	68
257	926	1,629	202	3,382	598	1,612	42,551	109,367	1,397	384	120,815	163,365	69
32	114	781	219	2,474	8,701	3,574	21,934	25,267	54	403	25,818	47,711	70
100	192	1,360	155	8,608	1,961	3,654	38,798	70,868	292	618	74,456	113,253	71
7,800	43,740	52,825	37,321	163,365	47,752	113,253		490,660	90,804	88,315	795,388		

plying their inputs will have to cut back. Moreover, asset values will change in many industries and consumer markets (for example, house prices will fall in some towns where automobile workers live, but rise in others where oil workers live).

It is also possible to trace *geographic* ripples. Thus, for example, changing New York real estate prices can affect the level of employment in San Francisco. As New York office prices rise, some firms will move their headquarters elsewhere, driving up office prices in an ever-widening circle around New York. Certain firms will relocate farther west, driving up rents there and inducing still other firms to shift, ultimately, to San Francisco. The effect is reinforced because rising house prices in New York will also induce some New York residents to move to other towns. That ripple will also encourage firms to locate westward, eventually reaching San Francisco.

Day in, day out, the economy accommodates an endless series of ripples among sectors and locales. Predicting them, tracing them, factoring each one out from the host of others that are occurring, these are standard economic skills. *Logic* is involved, in deciding which things are affected and in which directions, up or down. *Matters of degree* are also involved, in weighing which changes are largest and most important. For example, the price of cold-rolled steel has only a minor effect on the price of dog-racing tickets (and vice versa), but it significantly affects the price of automobiles. One learns to focus on the main changes, so as not to get lost among the lesser ones.

Limits on competitive efficiency

Remarkable though it is, the market system may not provide a completely satisfactory outcome. The Invisible Hand may falter, so that there are gaps in its results. These limits have long been recognized and studied. Sometimes their effects are minor, but often they have great force, causing large economic losses.

Market failures fall into five main classes, which we present in this section. In other chapters and in your own daily life, you will come to learn the many cases where they crop up and pose difficult choices for society, causing endless troubles that are major issues in the political process. One must understand them to judge their causes and cures.

External costs and benefits

Thus far, we have assumed that all prices and costs in the market are identical to the true social values and costs. If that holds true, then *prices* as used in consumer decisions will equal the ultimate social worths of the goods. The *costs* of private firms would also exactly reflect true social sacrifice.

But if external effects occur, then private values will no longer equal social values, and markets will cause inefficiency. For example, an automobile model may cost $5,000, including $950 worth of steel. At that price, it is bought and sold by many thousands of people. Yet, the making of that steel may have created water pollution and smoke that, per ton of steel made, costs other people $100 to clean up or to cure the medical harm. The steel companies do not have to pay those extra costs. Moreover, each car may release fumes that, when breathed, cost $50 worth of health damage to others. The ***external costs*** are $150 per car. The private parties—steel and automobile firms, and car drivers—act strictly on the private cost and price, which total $5,000 per car. Yet, the true ***social cost*** for producing and using each car is $5,150. The dollar difference may seem small here, but the logic, as illustrated in Figure 3, is clear.

Illustrations of Ripple Effects in the Economy

1. A series of military crises abroad causes U.S. officials to triple military spending for tanks, armored cars, and hand weapons. Consequently, production and prices rise for heavy metals, explosives, paint, motors, and other inputs for weapons. Costs also rise for other goods using those materials. Profits and stock prices of weapons companies rise, making their shareholders richer. The opposite happens for most other companies using those inputs, for higher costs squeeze their profits. Real estate values rise in the weapons-producing areas (parts of Texas, California, etc.).
2. The demand for large automobiles falls, while demand for small imported cars rises. Some 200,000 U.S. automobile workers are laid off. Production of automobile parts also falls, and workers in those markets also lose jobs. As the laid-off workers' incomes fall, the towns they live in suffer economic declines. Retail shops close, property values decline, and some of the population moves to look for better jobs. The stockholders of automobile and automobile parts firms undergo capital losses as the value of their shares falls.
3. The electronics industry (computers and related applications) encounters explosive growth. The wages for computer engineers and programmers rise sharply, attracting workers from other industries. That forces wages up in those other industries, raising their costs. Suppliers of electronics components also have a boom, as do those suppliers' suppliers. Their shareholders achieve capital gains. A series of adjustments toward electronics products, inputs, and workers' skills takes place.
4. An inefficient steel plant in Youngstown, Ohio, is closed down, putting 6,320 people out of work. Steel production goes down, and steel prices are slightly higher. Demand and prices of inputs fall off. Shipments of goods to and from the town are sharply cut, reducing jobs and incomes of railroad and trucking workers. Unemployed workers in Youngstown spend less for most items, reducing their quantities and prices. House prices drop, especially as people begin moving elsewhere. Local stores suffer and some fail. The cutbacks spread to other locales, whose products were used in Youngstown. Since virtually all activities in Youngstown are cut back, the ripple effects elsewhere cover a wide variety of products and locales.

Social and private values diverge whenever there are such external effects (often called "spillover effects"). ***Since only the "internal" costs—or market costs—matter to the private actors, they have no incentive to take external costs into account.*** Such external costs have been widespread, and despite the vast environmental

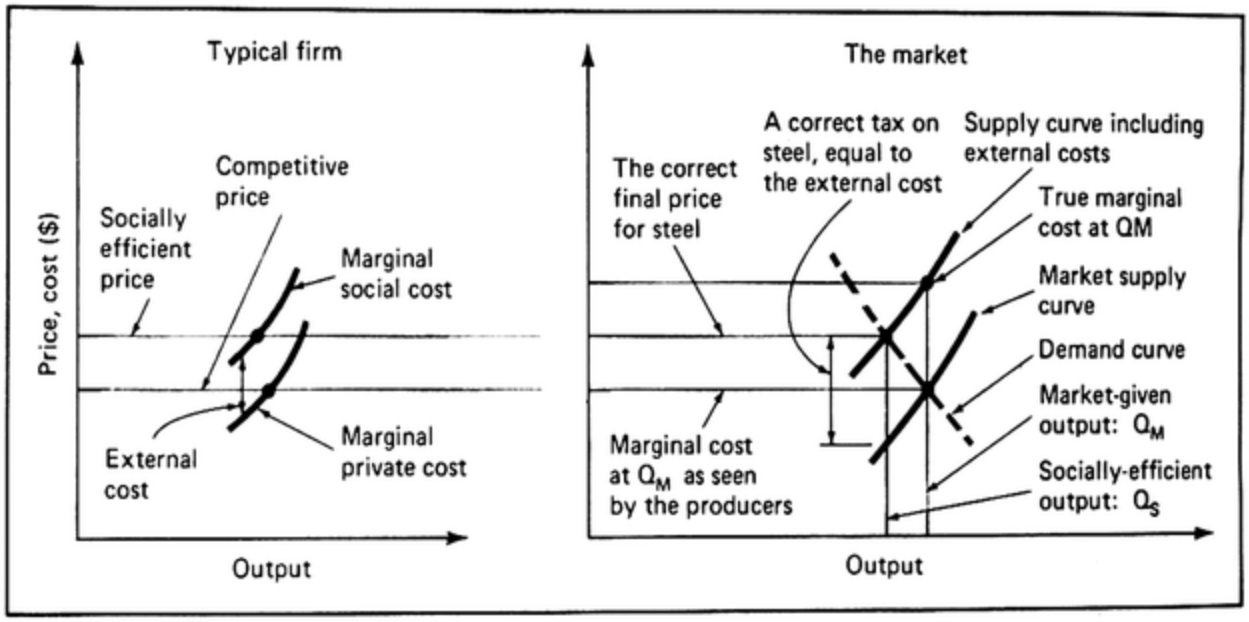


Figure 3 Adding external costs to private costs

Private costs are those incurred by the firm. They help determine how much the firm produces. But other costs may also be incurred, which are external to the firm's own interests. Pollution may occur, which is costly to others. This results in price (marginal benefit) being less than marginal cost at the market-given output. The marginal resources allocated to steel have opportunity cost in excess of the benefits they produce for steel buyers.

The full social costs—private costs plus external costs—are illustrated here. They may give a socially efficient level of output and price that differs from the market levels.

cleanup program in the United States since 1968, many such costs continue.

Whenever there are external effects, the competitive allocation will stray from the efficient patterns. In Figure 3's example, external costs will mean that the market output is too large and market price is too low, compared to the socially efficient levels. Such distortions from true social values will not be self-correcting, since the external costs are simply ignored by buyers and sellers acting through the market. If firms or car drivers need not pay for the harmful waste or smoke emissions they inflict on others, then the damages will occur. Allocation will be distorted from the socially efficient patterns.

The logic is simple but sound. You can now analyze and portray a class of deviations—or market failures—that competition will not correct. Indeed, competition will pressure firms to ignore external costs to survive.

Some familiar examples of both external costs and ***external benefits*** are given in Table 3. External costs may need to be reduced by social action: taxes, rules, limits, or others. Otherwise private choices will yield far too much of them to be efficient. Of course, because thousands of external effects are small, people let them pass as minor gains or irritants. But where external benefits are large—parks, schools, national defense, highways, lighthouses, courts, or police, for example—then the market process may fail altogether.

In such cases, special measures are needed to encourage production and consumption. Those extreme cases are called "social goods," for the benefits are large and widespread throughout society. Private markets will not provide them, because the sellers cannot confine the benefits just to the buyers. After all, anyone is free to go to a park, drive the streets, and enjoy the security of the nation's military

Table 3 Several examples of external costs and benefits

1. *External Costs*
   a. *In Production*
      Pollution of air and water by factory smoke and toxic wastes.
   b. *In Consumption*
      Fumes from automobiles; highway accidents, caused by reckless driving of automobiles.
      Loud music played in apartment buildings.
      Forest fires caused by careless campers.
      Nonsmokers' lung cancer, caused by smoke from others' cigarettes.
2. *External Benefits*
   a. *In Production*
      A lighthouse, whose beams are seen by all ships.
      Schooling, which provides an educated electorate and cultural support, as well as private skills.
   b. *In Consumption*
      Inoculation of one person against contagious disease gives added protection to others.
      A neighbor's beautiful garden and lawn, visible to others.

defenses. To sell them for private use would be absurd. Therefore, some or all of their cost must be met by public budgets, as we discuss in Chapters 18 and 20. Public goods are not just fringe matters. Some of them are central to a productive economy and a healthy society.

Distribution may be unfair

The Invisible Hand does not assure that income and wealth will be fairly distributed. ***Fairness*** is a matter separate from the allocation of resources. *Equity is, therefore, apart from efficiency.* Even if competition gives an utterly efficient allocation, distribution may be utterly unfair.

To illustrate the point, imagine a society in which a few rich aristocrats live in opulence while the mass of peasants suffers grinding poverty. Then suppose that an "industrial revolution" occurs, transforming the aristocrats into affluent business owners while the peasants are herded into factories as common laborers to toil at subsistence wages. The resulting economic outcome may be highly competitive and an efficient allocation of resources, but the unfair distribution between rich and poor may be no better than it was before. It may even be worse. Whether it be worse, better or the same, the Invisible Hand does not assure that the outcome will be fair.

Yet, the competitive outcome might be fair because wages are aligned with ***marginal productivity.*** Remember that in competitive markets each input (land, labor, capital, materials, etc.) is used up to the point where its marginal value product just equals its price. *Therefore, it seems, inputs are paid just what they are worth.* If the marginal value product of mill hands or assemblers is $3.50 per hour, then that is what they will be paid. That is all their labor is worth, and thus, perhaps, all they deserve to get. The same holds for interest as a payment to owners of capital, for the pay of presidents of companies, for the price of farmland—indeed, for any input into the competitive economy. ***Each input gets paid no less, and no more, than what it adds to the value of production at the margin.***

The logic seems impeccable, and it does reflect one general standard of fairness: "Rewards should fit contribution." Some 19th-century Manchester school economists carried the point to extremes, using it to defend low wages for workers and high profits for capital. Each input, they said, deserved exactly what it got in the capitalist system, no matter how unfair the disparities may have seemed by other standards, including need, equality, and the avoidance of destitution. Further, the economists said, attempts to alter the market outcomes would be self-defeating, for they would distort the system and cause more economic harm than good.

Most economists now do not accept that extreme view because people's skills and opportunities are often unfairly governed by their families' status, rather than by their own efforts and talents. Some children get much more help up the economic ladder—in schooling, job connections, and inheritances—than do other children. Furthermore, some children can be said to have an unfair advantage if they have by genetic chance inherited superior talents. Thus, the marginal revenue product of a ditchdigger or of a senior vice-president may be the outcome of an unfair process, with some people having greater privileges and help than others. Differing pay and wealth might be fair if every person had precisely fair chances all along, and if every person's rewards fitted his or her efforts.

Technological progress may not be at optimum rates. This possibility was presented in Chapter 16.

Cultural values are not necessarily provided

A competitive economy can be bleak. Competition itself is ultimately divisive, pitting people against one another to struggle endlessly in anonymous markets. Competition provides no mercy, no safety net, no relief from the pressure. Coldhearted, aggressive competitors will commonly do better economically than kind and sensitive people. Children may be sent into factories and mines to do degrading work for low wages—conditions that were widespread in the 19th century. The creative and performing arts (music, drama, painting, and others) will exist only if they can sell at a profit in the marketplace. Otherwise, cultural activities may be provided by nonprofit enterprises, if at all. Such enterprises are largely outside the private market system. Free-market capitalism may give progress and diversity or harshness and desolation; a favorable outcome is not assured.

Cultural health in a competitive system is, therefore, an open question. The heyday of competitive capitalism in Victorian Britain and the Gilded Age in the United States, for example, both inflicted cultural and social damage and brought certain economic and cultural gains. Indeed, the cultural effects—both harms and benefits—of a competitive system may ultimately outweigh the importance of its economic impacts. On the other side, one benefit of a competitive economy is that thorough competition makes political democracy healthier, by avoiding large concentrations of economic power.

All of these harms and benefits make it complex to judge the cultural results of a market economy. The main lesson here is that ***healthy social and cultural conditions are not guaranteed by a competitive economy; they may, instead, be limited or prevented by it, or be simply independent of it.***

Monopoly

Effective competition may be prevented in some or all markets, as we discussed at length in earlier chapters. ***Monopoly power*** may arise in two main ways; (1) There may be large economies of scale, (2) or one firm may simply seize a monopoly position.

Economies of scale If minimum efficient scale (the low point of the average cost curve) is reached at a small share of the market, then there is "natural competition." But if economies of scale prevail up to firm sizes that give a large share of the market, then the first firm to grow can undersell the others and take over much or all of the whole market. Such "natural monopoly" conditions make competition—and its spontaneous efficiency—impossible.

Monopolizing Even if scale economies are lacking, an aggressive firm may simply be able to buy out other firms or to use unfair

tactics to gain a monopoly. Indeed, every competitor in every market seeks to capture a higher market share for itself, but competition among these firms prevents any one firm from gaining dominance. Yet, sometimes a firm does attain a dominant position or even a pure monopoly. When this occurs—even if only a partial degree of monopoly is gained—the precise competitive results are distorted.

We need not recapitulate all of monopoly's effects here. Allocation is usually distorted away from efficient patterns to some degree, and innovation and equity may suffer. Against those probable losses, there may be gains from achieving lower costs because of economies of scale. The outcome is a matter of degree, of comparing the costs and benefits.

Natural resources

Competition may have special effects on ***natural resources.*** The resources of this planet fit on a spectrum from *renewable* (like crops and fresh air) to *nonrenewable* (like iron ore and oil), which, once used, are gone forever. ***Competitive markets can give the best economic use of these resources, conserving them, so that we use them neither too fast nor too slowly.*** That remarkable point—that free-market choices by resource owners tend to conserve many resources, not waste them—is presented in detail in Chapter 20.

Yet, there are three exceptions to this rule:

1. Common-property resources are available to more than one owner Whoever captures them can keep them. Fish and game animals are examples of such common-property resources. *The rate of use of these resources in a competitive setting will be too rapid.* Each user has an incentive to take the resource as fast as possible, because otherwise it will be taken by others. In total, all users will deplete the resource faster than is optimal. Hence, many great shoals are "overfished" under open competition. (That, too, is why the "Save the Whales" campaign is trying to prevent the effects of economic incentives on a common-property resource, the whales.) And the hunting of deer, elk, and other game animals has to be limited by seasons and licenses. Otherwise, the animals will be reduced in numbers and possibly become extinct.

2. External effects among resource users can distort their use (this is one class of external costs). For example, careless farming on a high slope can cause rapid runoffs of water, which then wash off valuable topsoil from other farms down below. The topsoil loss is external to the farmer on the upper slope. External effects can also affect air, water, and other resources, as Table 3 illustrated.

3. Permanence Free markets are particularly effective at solving short-run problems needing marginal (that is, small) adjustments. The same also holds true for most resources. The owners look ahead to use their resources efficiently now and for the foreseeable future. Market prices rise to reflect emerging scarcities.

But what of virtually permanent actions, where the effects last many centuries? Here, the competitive calculus may not give optimal results. For example, nuclear-power plants operate for only about 30 years and then must be closed down and sealed off for some 24,000 years—virtually forever by human time scales. If people could see the future better, they might build no such perpetual mausoleums, with their stores of nuclear material; or, instead, they might build more of them, if technology will make it easy to store the wastes or neutralize the radiation. There is simply no reliable current guide to those future choices. The next 24,000 years may bring unimaginably dif-

ferent conditions and needs, or they may settle down into grinding scarcity. At any rate, free-market choices may be too short-sighted in such long-run global matters.

These five main limits on the Invisible Hand could reduce the domain of true free-market efficiency to a few small zones, rather than the entire economy. Or perhaps the domain is nearly complete, at least in some economies at some times. Economists have found that the limits do affect some sectors more sharply than others. All of the limits lead to lively controversies in practice.

Summary

1. The study of general equilibrium combines all of the marginal conditions reached by consumers and firms in competitive markets. Together, these conditions define efficient allocation throughout the system. When it works well, the market system adjusts spontaneously toward the efficient allocation and, once in equilibrium, holds to it.
2. The key equimarginal condition for each firm is price equals marginal cost, at minimum average cost. When this occurs, all of the arrays of marginal utilities and marginal productivities are brought into line with their prices. Inputs are paid the value of their marginal products, and that may give a fair distribution.
3. Economic changes cause ripples to spread among sectors and regions. An input-output table helps one to visualize and trace these successive changes. Big and small changes can be traced through several or more rounds of adjustments.
4. If private values diverge from social values, then the Invisible Hand gives distorted results. There are five main causes of such market failures: (1) There may be external costs and benefits, which markets ignore. At the extreme, public goods will be entirely lacking. (2) The distribution of wealth and income may be unfair, even if allocation is efficient. (3) There may be a bleak culture and unhealthy social conditions, or cultural richness, or other results. The outcome is indeterminate. (4) Monopoly may occur, either via scale economies or simple monopolizing. (5) Natural resources may be used inefficiently in some situations. If resources are owned on a common-property basis, have external effects, or involve truly permanent changes, then competitive free-market choices may distort their use.

Key concepts

General equilibrium
Ripple effects
Efficient allocation
Marginal conditions
Invisible Hand
Market failures
External costs
Social cost
External benefits
Fairness
Marginal productivity
Monopoly power
Natural resources

Questions for review

1. *Is* every market related to all of the others? Try to give three instances of pairs of markets that are utterly *unrelated.* Then try to show that they are related, after all.

2. An earthquake shakes Los Angeles, knocking down hundreds of buildings and injuring 2,000 people. Trace the effects this might have on real estate prices near steel mills throughout the United States, on construction workers' wages in Los Angeles and Louisiana, and on the value of medical supply companies' common stocks.
3. "Price equals marginal cost is just a simple equation." Explain why it is also of profound importance.
4. Production of plastic pipes involves standard-shaped cost curves. Draw them, with average cost reaching a minimum of $1 per foot. Also draw supply and demand curves, intersecting at $1 per foot. The factories pour toxic wastes into nearby streams, whose cleanup costs are 50 cents per foot of pipe. Draw that extra cost in the cost-curves diagram and the supply-demand diagram. Show how the private prices and outputs depart from the socially efficient ones.
5. Define an efficient allocation under general equilibrium. Next, define a fair distribution, by any standard you prefer (or perhaps a combination of several standards). Will the efficient outcome be fair, either necessarily or by chance?
6. You are ambitious and hard-working, looking forward to a career that may eventually pay you $70,000 per year or more. Would that income be fair? Explain.
7. What good things do a decent society and rich culture provide? Which of these will a competitive economic system assure? Which of them might it not assure?
8. One day General Motors merges with Ford, Chrysler, and Volkswagen. In what directions would this probably change the prices and outputs of automobiles, steel, rubber, Detroit real estate, buses, and airline services?
9. The price of lobsters has gone from $4 to $13 per pound in recent years. How might competitive harvesting of this resource explain that rise?

18

Public Finance

As you read and study this chapter, you will learn:

- the nature of social goods and external effects
- how to analyze public spending decisions using cost-benefit analysis
- how to analyze the effects of taxes on distribution and incentives
- the main pattern of actual spending and taxation

When economics first took modern shape in the 19th century, it was called "political economy." The two-word name is fitting because two great parallel processes together shape the performance of the economic system. One is the *economic* process itself, with all of the technical conditions of allocation that you now know. The other is the *political* process, which proceeds continually in cities, states, and the federal government.

Economic groups try to advance their own interests in these two processes, both individually in the market system *and* collectively by trying to exert political pressure. One result of all these efforts is a set of government rules, spending programs, and taxing policies. They have widespread effects because government takes over 30 percent of total GNP in the United States. Some of those policies have been mentioned in Chapters 1–16, but an extensive analysis of their nature has awaited this chapter.

Public finance—the study of spending and taxation by governments—is an especially challenging microeconomic subject. Good policies can often be defined, as we show in the first section, using well-established economic concepts. Social cost, social goods, and cost-benefit analysis are crucial concepts that can clarify even the most intricate problems. They all involve the familiar comparison of marginal costs with marginal benefits.

With these concepts in place, we discuss in the second section how economists analyze three fundamental subjects: *Incidence:* Who really bears the burdens of taxes? *Incentives:* How do taxes and subsidies change people's working habits? And *distribution:* How might the structure of taxes affect the distribution of wealth and income?

Though optimal social policies can be analyzed abstractly, they may not be applied in practice. Indeed, governments' actual regulatory, spending, and taxing decisions may be thoroughly inefficient and misguided. No government is perfect, and many are abject failures. History gives many examples of flawed government from Nero to Richard Nixon, from Boss Tweed to many present city councils. In fact, many governmental follies are visible in the daily news reports. Few subjects are more hotly debated than pollution controls, military budgets, welfare spending, and taxes on property and income. The actual trends presented in the third section of this chapter are the outcomes of intense, incessant struggles at all levels of government.

Economic concepts of optimal public policies

Governments have economic effects through three main powers: rule making, spending, and taxation. *Rules* are of many types, covering a variety of conditions, such as monopoly power, labor unions, pollution, and hosts of lesser matters such as speed limits and housing safety codes. *Spending* and *taxes*, however, are the core topics of public finance. They deal with the *budgets* of governments: how they take in money from the populace, via taxes, and how they spend money in public programs that provide goods and services to the populace.

Economists traditionally assign three main purposes to the public sector: ***Efficiency,*** the use of spending and taxes to improve the allocation of resources; ***equity,*** to improve the fairness of the distribution of wealth and income; and ***stabilization,*** the use of the budget to reduce economic fluctuations. All three are often closely mingled in the rough and tumble of actual budgetary actions in Congress, in the state legislatures, and in every town and city.

Moreover, the three goals can be in conflict. An efficient program may cause unfairness and instability, or a fair redistribution of wealth by taxes may hurt incentives for efficiency, and so on. These conflicts among goals may be important in some cases but minor in others. Over large areas of choice, the goals may be in harmony rather than in conflict.

In any event, economists try to keep the three goals clear as concepts. Each is valuable, and each can be served by good fiscal policies.

Social goods

Government policies are necessary because private markets often fail to reach optimal conditions. The main causes of such *market failure* are social goods, external effects, inequitable distribution, common-property resources, and monopoly. Table 1 gives examples of these market failures. Ideally, government policies will correct these problems by applying the principles

Table 1 ***Categories of market failures***

Categories	Examples
1. *External effects:* External costs and benefits are ignored by private decision makers.	
a. *External cost:* Private choices cause output to be too large, compared to the optimum.	Pollution of the air and water by toxic wastes, smoke, etc.
b. *External benefit:* Private choices cause output to be too small, compared to the optimum.	Too small levels of: an educated electorate and worker skills (that could be provided by public schools); national defense; parks; police; lighthouses; public health actions against contagious disease
2. *Inequitable distribution*	**Departures from criteria of fairness; e.g., large payments to nonproductive people, low payments to hard-working and/or needy people**
3. *Open access to common-property resources:* Private actions cause excessive depletions of the resource.	Oceanic fish, lobsters, oysters, salmon, whales, etc.
4. *Monopoly:* Output is restricted; price is raised; innovation is retarded; etc.	**Postal service; water and sewage; urban transit systems; courts; fire fighters**
5. *Others*	
An unfair distribution of wealth and income	Extremes of wealth and poverty
An exclusion of some people from adequate health care, insurance or other needs	"Uninsurable" people; sick people unable to afford minimum health care
Hazardous jobs and products	Workplace fatalities; dangerous toys; carcinogenic products
Other hazards that a "decent society" protects its citizens against	Mass unemployment
"Nonmarket" values: There is no price tag possible, so the market will not produce them, although society may value them.	Justice; freedom from degrading and dangerous activities

of efficient resource allocation to the public sector.

A ***social good*** differs from a private good in one crucial respect: Consumption by one person does not reduce the supply available to others. ***Social goods are nonexclusive (or nonrival) in consumption. Private goods are exclusive:*** If Jones buys and uses a loaf of bread, chair, automobile, or house, each item is unavailable to everyone else. But a social good (such as a park, a road, and fire protection) is simultaneously enjoyed by many or even all citizens.

A classic instance of a social good is national defense. The military defense of the country provides a unified service. That service cannot be sold piecemeal to

individuals to fit their personal desires. Instead, all citizens share alike in the resulting level of national security. A private market, by contrast, might generate various private armies for certain wealthy groups, but there would be no efficient country-wide system, and thus no true "national defense." The same logic applies, in lesser degree, to such things as public parks, police and fire protection, and public roads. Private markets would supply little or none of them.

Because a social good is consumed nonexclusively, many or all users of the good can be "free riders." Once a social good is provided, more people can use it. All users will, in fact, wish to be free riders, to use it without paying. For example, people will expect to have free access to a park or to police protection, once they are in operation. But when all people are free riders, none of them makes any payment for the social good. There is no economic expression of demand for the good.

Therefore, because social goods will not be supplied by private markets at all (or they will be undersupplied), they are a social (governmental) responsibility. The government must then decide (1) *whether* to supply the good at all, and (2) exactly *how much* of it to provide. The question of *whether* is important: Many possible social goods should not be provided at all. But the question of *how much* is the more complex issue, for which economists have prepared the following method.

For social goods, demand is summed vertically To fit the unique character of social goods, economists have developed a distinctive analysis of the demand for them. People still have their private preferences for these goods, in choosing between them and all others. Thus, people's willingness to pay for parks, schooling, and national security can be represented by individual demand curves. These demand curves show how much people would each be willing to pay (if they were required to pay) for varying amounts of each public good. Figure 1 illustrates the case of police protection in a city. People's demands may vary, as shown, rather than be rigidly identical. Thus, Alpha's demand is much above Beta's, because Alpha has (1) stronger preferences for police services and/or (2) more money to pay for them.

But these demands are not summed up horizontally, as individual demands for private goods are (recall Chapter 6). Instead, they are summed up *vertically*, as shown in Figure 1, because the good cannot be consumed separately. You do not take a public park home for your own use, the way you do a tomato or a pair of shoes. The park is open to all, whether you use it or not. Therefore, the desired quantity of the public good, say, a park or road system, has to be the same for all. **Only the *summation of what people are willing to pay* gives the basis for judging how much of the public good should be provided.** That summation is done vertically, as shown in Figure 1. Then the efficient amount is given by the intersection of the demand curve with the supply curve (Point *A*). At that point, marginal benefits just equal marginal cost. At a higher quantity, marginal costs would exceed marginal benefits; the extra output costs more than it is worth.

The supply curve's rising slope reflects the added costs needed to obtain larger amounts of the good. That is perfectly conventional.

Ideally, public officials would discern people's demands, add them up, lay on the supply curve, locate the intersection of total demand and supply, and thus neatly fix the optimal level of each public good. The right amount of spending for defense, public parks, schools and public universities, welfare programs, and the thousands of

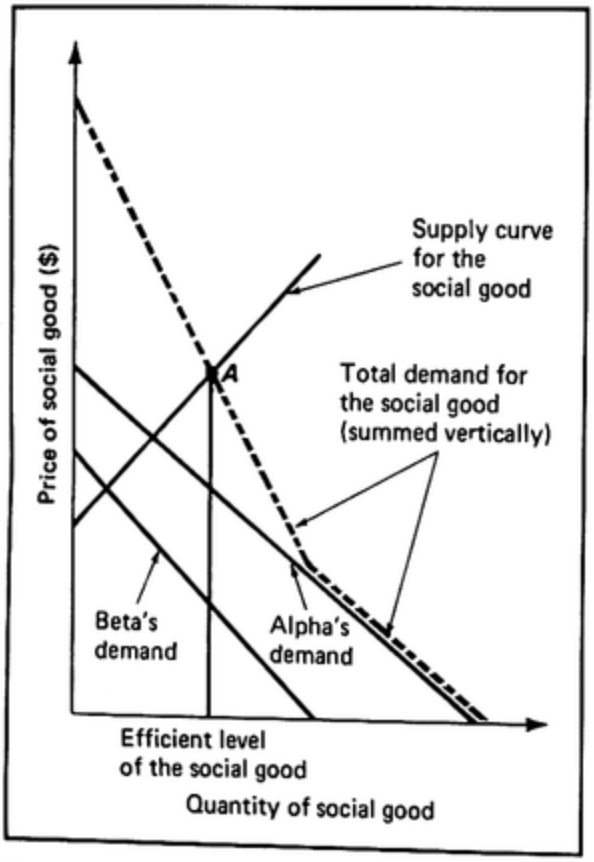


Figure 1 Summing the individual curves *vertically* for a public good

Public goods are available to more than one person, perhaps to many, in contrast to private goods. Therefore a vertical summing of individual demand curves is logical, as shown for the two people in this illustration. The question is not how much they would each buy at each price, but rather *how much the two persons together would be willing to pay for each level of public goods.*

other public goods could be decided clearly and precisely. Efficient conditions could be found and achieved.

The ideal process is not possible in practice, for two main reasons. *First, people are likely to misrepresent their true demand.* The public good is nonexclusive: If some people can pay for the good, other people can enjoy it for free. This is the free rider problem, as we noted earlier. Everyone would like to be a free rider, claiming to have little demand for the public good, but using it nevertheless when it is supplied. If people know their statements about their demands will commit them to pay a certain price, they will understate their true levels of demand. Other people, they hope, will pay for the public good. But, of course, if all people claim that their demands are low, then the public good may not be produced at all.

Conversely, if citizens' answers do not affect how much they have to pay, they will be induced to overstate vastly the value of the public good to them. Yes, they will say, we should have enormous parks, fine schools, superb roads, and so on (so long as I'm not paying for them).

Both approaches to discovering demand are therefore wrong; One understates demand, the other overstates it. For ordinary private goods, by contrast, people's actions in markets directly express their willingness to pay.

Second, people may be unsure about their preferences, even if they try to give honest answers. They are accustomed to deciding about ordinary goods' prices and purchases scores of times a day. But they are rarely asked to set a money price on parks, schools, national defense, roads, and the like. People want them, in some degree, but they will often be unable to express consistent demand curves.

Altogether, social goods routinely stir exaggerated claims about people's needs and wishes. But no neat metrical process, involving economists taking polls, drawing demand curves, and finding the demand-supply intersections, can be used to decide their optimal levels. Instead, there is a political process to thrash out the choices, in a loose sequence of public debate and compromises among interest groups. The outcomes are inherently controversial, for (1) various groups have opposing interests, (2) people are often rewarded for misrepresenting their attitudes, and (3) the underlying true conditions are often unsure and changeable.

External effects

Between private goods and pure social goods lies a band of goods that are private

but also have ***external effects.*** Such goods are partly social in nature, because the external effects create social values that the market process will ignore.

A simple instance of the divergence was given in Chapter 17: Pollution caused in producing steel is not a cost that the steel company must pay, but it results in real cost to society. Though the company can and does ignore the cost of pollution, society cannot. Public policies are often applied to allow for such external effects; pollution-control programs are one example. ***But the policies are appropriate only if their costs are less than the external costs that they are designed to prevent.*** If the steel company pollution's external costs are $20 million, then a pollution control program costing more than $20 million would be wasteful. In short, an external effect invites corrective public action but does not require it; the issue is one of cost.

External effects occur whenever production has repercussions beyond the seller and buyer of the good. The externalities can be either ***costs*** or ***benefits,*** as illustrated in Table 1. They are often called *social* costs and benefits to separate them from the strictly private costs and benefits that are covered by private market choices.

External costs, such as pollution, have often been severe in industrial areas. London, Pittsburgh, and scores of other cities were shrouded in thick and unhealthy smog. Despite cleanup efforts since 1950, air pollution persists in places as diverse as Gary, Indiana, Los Angeles, and New York City. Chemical wastes at hundreds of dumping sites are another large problem. There are many other external costs, including noise, traffic hazards, and even such subtle effects as those a large new building causes when it blocks the sunlight and views of nearby buildings.

External benefits arise when a good provides value to people who do not have to pay for it. Thus, a beautiful house and lawn or a handsome office building gives pleasure to people passing by. The owners pay for them, but the benefits are also available to others. External benefits are often less obvious than costs, but they can be important.

These external effects range from small to large, compared to the size of the private benefits and costs of goods. Ideally, the external element could be analyzed separately, by adding up the external benefits vertically. The private demands, meanwhile, would be added horizontally to form the market demand curve. In practice, however, such a two-part arrangement is difficult to apply.

Now we turn to the general methods for setting the levels of public expenditure for public goods or external effects.

Public expenditure: Cost-benefit analysis

Public expenditure requires choices among a large array of possible projects, whose total is far more than the public resources available. To allocate those scarce resources efficiently, economists rely mainly on ***cost-benefit analysis.*** Before presenting that set of tools, we need to show how spending is to be allocated among alternative projects and between social and private goods.

Allocating spending among goods Suppose that there are two social goods, *A* and *B* (they might be military and civilian programs, highways and welfare, schools and police, or any other pair). Their benefits are known, as illustrated in Figure 2 by curves showing the marginal benefits of successive dollars spent on *A* and *B*. Those curves slope down, to reflect the diminishing marginal utility of these goods. Dollars

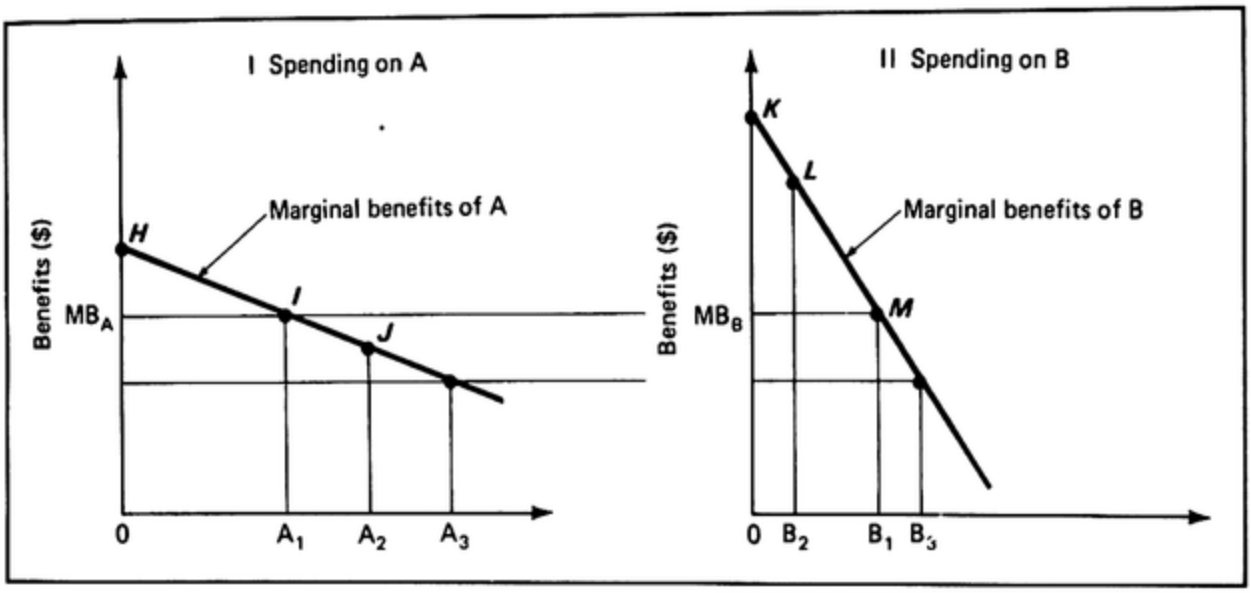


Figure 2 Allocating funds between alternatives

Let *A* and *B* be alternative social goods. The marginal benefits of each alternative are shown in Panels I and II; the downward slopes reflect declining marginal benefits.

An efficient choice sets spending so that the marginal benefits of spending on the two goods are equal. That condition is reached at A_1 and B_1. Total benefits ($OHIA_1$ plus $OKMB_1$) are maximized. For larger total spending, both levels increase, but the equality of marginal benefits is still needed, as at A_2 and B_2.

The same logic applies to allocations between public and private spending. If *A* were public and *B* were private goods, solutions such as A_1B_1 or A_2B_2 would be efficient; A_2 and B_2 would be inefficient.

not spent on *A* can be spent on *B*; therefore, the opportunity cost of a dollar spent on *A* is the marginal benefit lost by not spending it on *B*.

An efficient choice will allocate spending between *A* and *B*, so that the marginal benefits from each social good are equal. Thus, the levels A_1 and B_1 will be set, as shown, with $MB_A = MB_B$. The sum of total benefits is measured by the areas $OHIA_1$ for Good *A* and $OKMB_1$ for Good *B*. Those total benefits from both goods will be maximized when A_1 and B_1 are chosen. Any other selection (as at A_2 and B_2) will reduce the total benefits. The benefits area, A_1IJA_2, gained by spending A_1-A_2 more on Good *A* is much smaller than the benefits area, B_2LMB_1, lost by spending B_2-B_1 less on Good *B*.

If total spending is changed, then the amounts spent on *A* and *B* will also change, but the marginal benefits must still be equal. If spending increases sharply, the new efficient equilibrium might be at A_3 and B_3; the marginal benefits are lower than MB_A and MB_B, but they are equal.

We have shown how a given spending level is allocated. What if the spending level is variable? The analysis also shows what level should be chosen. The guiding principle is that *the opportunity cost of public spending is the benefit lost by taking resources away from private projects in the economy.* The goal is now to maximize the total of *public and private benefits.* The solution involves allocating dollars until the marginal public and private benefits are equal. This corresponds to the choice among social goods, and Figure 2 can illustrate the outcome. But now Good *A* is assumed to be private goods, while Good *B* is public goods. For any given level of total economic resources, the efficient allocation is reached when marginal

private benefits from A equal marginal social benefits from B: A_1B_1 or A_3B_3 would satisfy that criterion.

Cost-benefit analysis for specific projects Cost-benefit analysis is a formal method for deciding the efficient size of public projects. It was developed in the 1930s to analyze projects to dredge waterways and to build dams for flood control. Refined in the 1950s, it became in the 1960s the formal basis for all federal budget appraisals (under the name of "planning-programming-budgeting analysis," or PPB). The enthusiasm for it has receded, yet cost-benefit analysis is still a valid basis for framing the issues. Its core is sound and simple, though there are difficulties that keep it from giving definitive solutions. We present the technique and the problems.

THE CATEGORIES OF COSTS AND BENEFITS The first step is to define and measure all of the costs and benefits of a project. These values are of the following main categories: *tangible or intangible* and *direct or indirect*. *Tangible* benefits and costs can be measured in the market, using dollar figures from actual transactions. *Intangible* values are real but not readily assessed in money terms. For example, education raises students' future earnings, which is a tangible benefit; it also gives the intangible benefits of greater understanding and a richer life.

Direct costs and benefits are closely related to the project's main purpose, while *indirect* values are by-products or tangential. The distinction is meaningful, even though it is hard to define rigorously. For example, a university may directly provide education, but it indirectly adds to research and technological progress, and it also increases the prosperity and the quality of life of the town where it is located.

These categories are illustrated in Table 2 by costs and benefits for irrigation projects, education, and space research. Typically, some values can be estimated, but others can only be surmised. **Still, *the crucial first step is to include all significant kinds of costs and benefits, so that the comparison is complete, even if imprecise.*** Omission of any main class of costs or benefits can bias the evaluation and lead to major errors in policies. For example, the foregone earnings of college students are a large opportunity cost; ignoring it would encourage an overexpansion of public campuses, absorbing some students whose efficient choice is to take paying jobs. Conversely, one might ignore the research and new technology created by university faculty. That would understate total benefits and cause too little to be spent on education.

THE ANALYSIS The purpose is to maximize the net social benefit from each project. That result occurs when marginal benefits equal marginal costs. Stated simply:

Net social benefit
= (Total social benefit) - (Total social cost).

When net social benefit is maximized, spending is at the level where

Marginal social benefit = Marginal social cost.

The result is illustrated in Figure 3 for any typical project. All costs and benefits are included, even though some may be estimated rather than precisely known. The efficient level is OA, where marginal costs and benefits are both at the level AE. Total costs are $OBEA$. Total benefits include consumers' surplus in the amount BCE, so that the total benefits $OCEA$ exceed the total costs $OBEA$.

Alternative levels are inefficient. Level A_1 is too small, for it fails to achieve the net benefit DEF. Level A_2 is too high, for it causes the loss of EGH in net benefits. Note how the logic parallels the marginal cost-benefit choices that have recurred

Table 2 ***Illustrations of the main categories of costs and benefits***

Costs		Benefits
IRRIGATION PROJECT		
Tangible: Direct:	Cost of pipes, channels, and other facilities	Increase in farm output
Indirect:	Diversion of water from other uses	Reduction in soil erosion
Intangible: Direct:	Loss of wilderness area	Beautification of the area
Indirect:	Destruction of wildlife	Preservation of rural society
EDUCATION EXPENDITURE		
Tangible: Direct:	Teachers' salaries, cost of campus facilities, books, and related items	Increase in students' future earnings
Indirect		Reduced costs of preventing crime, because added skills reduce crime rates
Intangible: Direct:	Foregone leisure time	A richer life, with greater understanding
Indirect		A more intelligent electorate, greater political stability
SPACE RESEARCH		
Tangible: Direct:	Costs of inputs (workers, equipment, fuel, etc.)	As yet unknown
Indirect:	Diversion of talent and research from earth-based problems	The generation of new technology (discovering resources, flight, communications, etc.)
Intangible: Direct:	Pollution of space	Fascination of discovery; knowledge of the universe
Indirect		Gain in world prestige; military advantages

Source: Adapted from Table 8–2 in R. A. Musgrave and P. B. Musgrave, *Public Finance in Theory and Practice* 3rd. ed. (New York: McGraw-Hill, 1980).©McGraw-Hill Book Company, 1980. By permission.

throughout the discussion of microeconomics in this book.

Matters of degree are also important, as Figure 4 illustrates. If costs are underestimated, then the true cost curve lies above the apparent cost curve. The correct level is at A_3, less than half of A. If instead, benefits are overestimated, then the true benefits curve lies below, and the correct level, A_4, is even lower. Indeed, virtually none of this project should be done. Such contrasts occur frequently in practical cases, including irrigation, education, and space research. The project's advocates foresee large benefits and small costs; its critics expect costs to be large and benefits small. Whether the project should be large or terminated, then, turns on the correct amounts in the cost-benefit calculation.

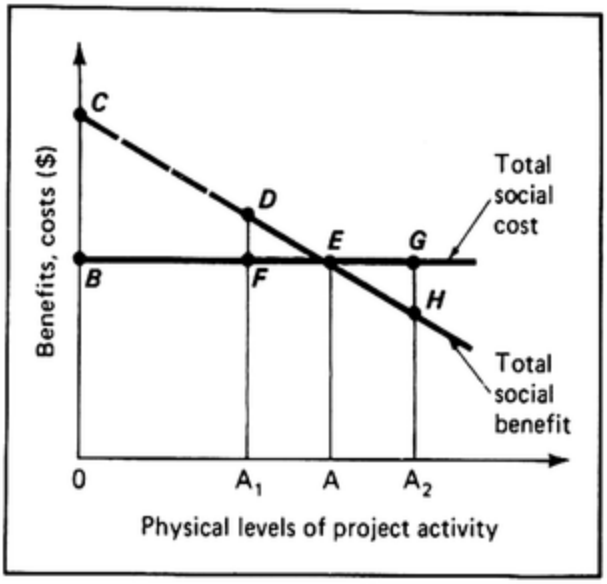


Figure 3 Simple cost-benefit analysis

The efficient level for the project is *A*, where both the marginal benefits and the marginal costs of the project are the amount *AE*. Level A_1 is too small because it would forgo the net benefits *FDE*. Level A_2 is too high because it incurs the unnecessary net loss of benefits shown by *EGH*.

DISCOUNTING The costs and benefits occur over time, rather than instantly. Though most of the costs are immediate, the benefits often stretch out over many years. For example, the construction of a public university classroom building requires current spending, but the building's benefits will only be realized during many future decades.

Such future values must be discounted because future benefits are less valuable than current benefits. The same is true of costs. The future values must be converted to *present values*. The discounting is similar to the discounting of private values done by private firms (in Chapter 7).

The discounted value depends on (1) the duration of the time interval and (2) the discount rate. For example, a $1 million value 10 years hence discounted at 5 percent has a present value of $613,900; for a 20-year interval, it is $376,900; for 20 years discounted at 10 percent, the present value is only $148,600. The most common error is to discount future benefits at too low a rate. That can cause a project's benefits to be sharply overstated and encourage too much spending on the project.

The correct discount rate is a matter of debate, but the consensus is that it should be at the interest rates prevailing in the private economy, with possibly an

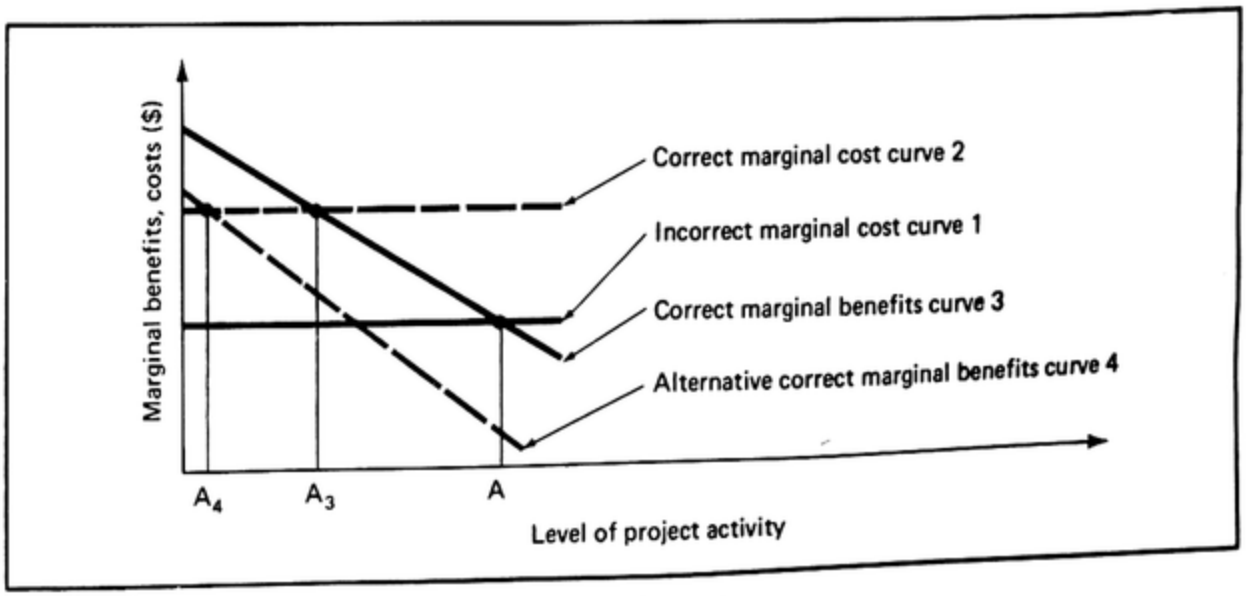


Figure 4 How wrong estimates can affect the outcomes

If officials believe that Curves 1 and 3 are correct, then amount *A* appears to be efficient. But if Curve 2 is correct (rather than 1), then the efficient level is much smaller, at A_3. If, moreover, Curve 4 is the correct representation of benefits (rather than Curve 3), then the efficient choice is at the low level of A_4.

adjustment to reflect social values. Private interest rates are usually much higher than those at which governments can borrow funds. Private rates would, therefore, apply a tighter standard to public projects, sharply reducing the efficient level of spending.

For example, for many years the Army Corps of Engineers used the interest rate on government borrowings (then about 3–5 percent) as its discount rate in justifying the building of hundreds of dams. But those projects took capital from the private sector, where it could have earned perhaps 10–12 percent on average. Since the opportunity cost of the capital was the forgone 10–12 percent rate, the Corps should have used a discount rate in that range. That would have shown that scores of the dams were an inefficient use of resources and should not have been built.

WHO SHOULD PAY: GENERAL TAXPAYERS OR SPECIFIC USERS? Even if the efficient level is defined, there is still the question of who should pay for the public service. Funds can come from *general tax revenues;* or the *users* can be charged for some or all of the costs; or both sources can be used, in some ratio.

General tax funds are the best source to use when the benefits are widely spread. Such universal social goods include military defense, police and fire protection, public health programs, and schools. The costs are shared because the benefits are shared.

At the other extreme, because ***many social goods are used only by specific groups of people, those groups ought to bear the financial burden.*** Examples are airports, water supply, harbors, many state and national parks, and certain roads. "User fees" or "earmarked taxes" are often charged to cover part or all of the costs. The fees may be general (e.g., a gasoline tax that pays for building and maintaining roads) or specific (e.g., a park permit or a toll charged for crossing a bridge).

Naturally, each user group tries to get its social good included in the general budget to minimize its own payments. Often, as a result of impassioned argument, deceptive data, or political muscle, such groups get their way. Therefore, many narrowly used public programs are financed out of general tax revenues, even though they benefit small groups of affluent people. Only when the recipients are poor and unable to pay is there a good reason for drawing on the general budget.

In practice, public funding is a patchwork of many varieties, with many departures from what seems appropriate. Moreover, the actual levels of many programs are also sharply inefficient. The political process is not perfect, and it often gives inefficient and unfair outcomes. Details on major public policies will be given in the third main section of this chapter and in Chapters 19 and 20.

Alternatives to public spending and taxes

Often a social need can be provided by means other than public spending and taxes. Four such alternatives follow.

Rules and fines rather than spending or taxes Social action can often rely on the setting of rules rather than on direct spending and taxes. For example, pollution can be reduced by declaring certain levels of it to be illegal, with fines imposed on violators. Firms would then adjust their profit-maximizing choices toward lower pollution levels. If, instead, the government gave money to the firms for cleaning up, or provided tax incentives to them, the dollar cost to the public budget might be much greater.

Setting such limits correctly is an important task, for which economics provides

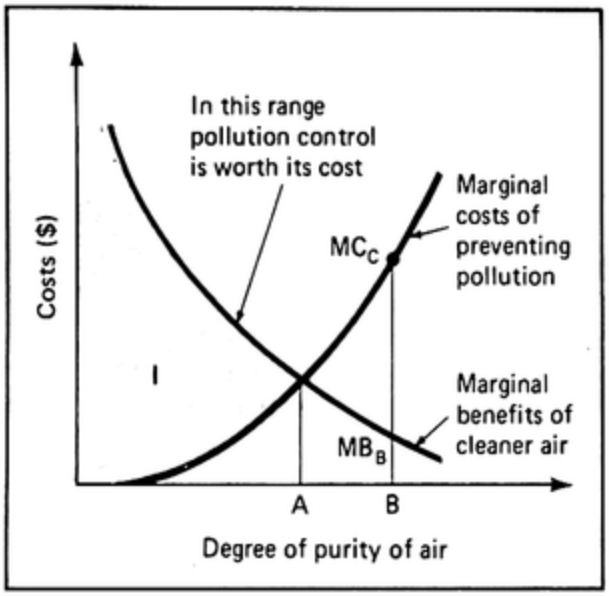


Figure 5 Costs and benefits of preventing air pollution

One solid line shows the marginal social benefits of clean air. It is high at the left, where pollution is intense; but as pollution recedes (toward the right), the *marginal benefits* of cleaner air decline to low levels. The *marginal costs* of preventing pollution follow the opposite pattern: They rise sharply as efforts are made to filter out the last few degrees of pollution (to the right).

Given these curves, the efficient level of cleanliness of air is at Point *A*, where marginal cost just equals marginal benefit. At a higher pollution level, the marginal damages of pollution are above the cost it would take to prevent them. The shaded area I is the net gain from reducing pollution to Point *A*. To get the air cleaner than it is at Point *A* would cost more than the damage it would prevent. At Point *B*, for example, the marginal cost of clean air (MC_c) is far above the marginal social benefits of the clean air (MB_s).

a clear analysis, as illustrated in Figure 5. "How much pollution" is a matter of degree, with higher levels imposing ever-higher marginal costs on society. But the costs of preventing pollution are also real, and they rise too, even at the margin, as pollution is reduced toward zero. Thus when automobile emissions were first attacked by law after 1965, it was relatively easy to prevent about half of the fumes and particles. Small adjustments in engine design were cheap and effective. But stopping the last 20 percent of harmful emissions (at the right-hand end of Figure 5) is far more costly, requiring major design changes and costly apparatus. The same is true of most factory smoke and chemical wastes.

The economic task is to compare marginal benefits and costs, to stop pollution only as far as it is worthwhile to do so. That balance is reached at Point *A* in Figure 5, where the marginal benefit of preventing that last bit of pollution equals the marginal cost of doing so. Reducing pollution even further would require using resources that cost more than the extra benefit they give.

Some students resist the idea that the pollution below *A* is economically acceptable: All pollution is bad, they say. Possibly, but when the marginal cost of stopping it is much higher than the extra benefit, the economist can only conclude that stamping out every trace of pollution makes no economic sense. To do so is too costly. Of course, the locations of the cost and benefit functions are often debatable.

Differing estimates of relative costs and benefits imply different policy actions. There is little wonder that the issue is controversial. For example, officials in the Reagan administration have argued that the marginal costs of pollution control are understated and the marginal benefits are exaggerated.

Another problem arises *when the costs of action are borne by different groups of people from those who gain the benefits.* For example, air pollution may hurt all people living in New York City, but the costs of cleaning it up might fall on factory owners and truck owners living elsewhere. Efficiency requires that a cleanup occur up to the correct margin, shown by Point *A* in Figure 5, but the factory and truck owners will resist the program. New Yorkers will advocate it, naturally. In such cases, there are political difficulties in achieving the efficient result. Yet, the economist's main task is still to define that efficient outcome.

These issues are perfectly general, applying to all external effects. A balance is usually needed between the marginal benefits and costs of the public action. But

now comes a subtle point: The costs of correction are equally valid, whether they are borne by the government budget or by private firms. If stopping factory smoke requires expensive new furnaces, that cost matters whether the companies or the government pays for it. Therefore, the cost curve in Figure 5—as well as Point *A*, the efficient level—applies regardless of who pays the costs.

Once Point *A* is known, the government can require companies to meet it. If enforced, that simple law gets the desired result without requiring public spending. The government could, instead, try methods that absorb public funds: To reach Point *A*, it could guarantee loans, provide tax breaks, or even buy new furnaces for the factories. But the government can minimize the taxpayers' burden by simply imposing the efficient rules.

Private sources Often the public service can in fact be supplied by private firms who see the need, design the service, and find a way to sell it. Thus, private "health maintenance organizations" arising since 1970 may obviate the need for more expensive public medical programs. Even where a social element remains, it can often be met by a limited subsidy rather than by a total public subsidy. Thus, public universities usually draw only part of their funds from governments. They rely on private tuition and fees to cover the private values that they provide to students.

Insurance pooling Risks of accident, disease, loss of work, and other calamities can be covered by government programs. But often private insurance schemes can be developed to cover most or all of such risks. As long as the pooling permits the system as a whole to meet its costs, coverage can be extended to include even people with high risk levels.

Charity and other not-for-profit suppliers When private firms fail to provide goods that have external benefits, government action may not be needed if special not-for-profit firms fill the gap. Until recent times, churches and charities provided most social assistance to the poor and helpless. Though often meager and demeaning, such charity support was frequently important. Even with the enlargement of public programs in recent decades, many not-for-profit firms still exist, including hospitals, cooperatives, the Red Cross, YMCA-YWCA, cultural organizations, and United Way groups.

These so-called third sector groups now make up a substantial part of the economy. Moreover, they cover some of the most distinctive social activities in our culture. In many sectors, they provide services that make public spending unnecessary.

Altogether, these four categories cover many social needs in ways that avoid public spending. The art of public finance lies in *minimizing* public spending while attaining efficient and fair social results.

Categories of spending and taxes

We conclude the first main section of this chapter by setting forth the standard categories of public spending and taxation. They provide background for analyzing tax impacts in the second main section and the major patterns of public finance in the third main section.

Spending Governments spend money on thousands of different items, from battleships to teachers' salaries to welfare payments. These items divide into two main categories:

PURCHASES Governments buy labor, products, and services. The money is spent to obtain costly items. Thus, people work for

governments and receive salaries. Companies sell governments such products as police cars, foods, aircraft, garbage trucks, desks, and all the rest. Governments buy such services as the building of schools and the paving of highways.

TRANSFER PAYMENTS This category is distinct from purchases, because ***transfer payments*** simply provide money to people who are in certain categories. The people do not supply any work, product, or service in return. To qualify for transfer payments, people must be poor, elderly, sick, or in some other category specified by the programs. The payments simply transfer funds from taxpayers to recipients. Of course, the recipients then spend the funds on goods and services that they choose.

Purchases provide governments with the supplies they need to function. Transfer payments, on the other hand, are meant to provide needy people with the funds they need to function—to spend as they choose.

Taxes A tax takes money from a person or organization by a government. ***Taxes*** can be imposed at many points in the economic process, such as on sales, income, property, or imports. Actual taxes present a wide variety. Yet, all taxes divide into two basic categories.

PERSONAL TAXES are based on the taxpayer's personal ability to pay. The leading example is the personal income tax, levied on the yearly flow of income. Most personal taxes are *direct* taxes, coming directly out of personal income.

IN REM TAXES ("taxes on things") are levied on objects or activities, such as sales, purchases, transfers of property, or the holding of property itself. These taxes are not based on the taxpayer's personal ability to pay. Most *in rem* taxes are *indirect* taxes, being imposed at various points in the economy rather than directly on the people who will finally bear the burden of the tax.

For example, a sales tax is an *in rem*, indirect tax. As we showed in Chapter 5, its burden depends on the relative inelasticities of demand and supply. An income tax is direct, coming straight from the pocket of the person who pays it.

Taxes: Impacts on distribution and incentives

We now discuss three effects of taxes: (1) on who bears the burden of the tax, (2) on incentives to work, and (3) on distribution. Economists have developed standard analyses of each topic, using supply, demand, and other concepts that you have now mastered. Their logic is straightforward, even though the matters of degree are often complex.

Incidence: Analyzing who bears the burden of taxes

Each tax dollar must come ultimately from someone. But that burden is often difficult to trace. For example, though you may pay $4 in sales tax when you buy $100 worth of textbooks, the $4 may ultimately be paid not by you but by others: perhaps the bookstore, or the publishing companies, or still others. For another example, private landlords usually pay real estate taxes on their apartment buildings. Do student renters end up paying those taxes indirectly?

These are matters of ***incidence***, the real burden of taxes. Incidence matters because tax burdens can be heavy, and where they fall is often highly uncertain. Tax burdens are often *shifted* from one group to another. Because personal income and wealth taxes cannot be shifted very much, their incidence falls directly on the payer. But since sales taxes and taxes on busi-

nesses may be shifted extensively, economists have studied their incidence in great detail.

In Chapter 5 we presented the main technique for analyzing the incidence of a sales tax. As you may recall, the tax is added to the supply curve of the good, as was shown in Figure 7 of Chapter 5. This shifts the market equilibrium from Point *A* to Point *B*. Output shrinks from 100 to 80. As for price, the consumer now pays $35 for each unit, which is more than the original $30 price before the sales tax was imposed. But the seller only receives $25 per unit, after turning over the $10 in tax to the government.

Therefore, in this case, the burden of the tax falls equally on the sellers and the buyers. Buyers pay $5 more per unit; sellers receive $5 less. This illustration was designed to show that the burden can fall on both groups. But the perfectly equal division it depicts is a fluke; the burdens of most taxes will be unevenly divided between sellers and buyers, perhaps even falling entirely on one group or the other.

What determines the outcome? How can the division of the burden be discovered? We presented the analysis of incidence in detail in Chapter 5. The basic answer is: *The burden depends on the relative elasticities of demand and supply.*

At the extreme, if either demand or supply is perfectly inelastic, then buyers or sellers, respectively, will bear the entire burden of a sales tax. More generally, the relative incidence of burdens reflects the relative inelasticities. Both demand and supply may be inelastic or elastic, but the burden will fall more heavily on the side that is *relatively less elastic* than the other.

This fact has long been recognized by tax officials, who have always sought inelastic items to tax. Inelasticity means, by definition, that there is little response of quantities. Therefore, the tax yield will be relatively high.

Examples of heavily taxed items with inelastic demand are such necessities as gasoline today and salt and spices in the Middle Ages, and such habit-forming items as liquor and tobacco. Though governments often claim to be taxing liquor and tobacco to discourage drinking and smoking, a primary reason for such taxes is that demand for these things is inelastic. As for inelastic supply, land is a prime instance, because it cannot be moved. Therefore, it is not surprising that most cities rely heavily on real estate taxes on land values.

Incidence also affects the equity of taxes. If the burden falls on people with low incomes, the tax may be unfair. One intriguing example is real estate taxes on apartments. Does their burden fall mainly on renters or on landlords? The tax is usually reflected in higher rents, so that renters pay at least part of the tax. But the exact shares depend on relative elasticities.

Taxes also affect efficiency by changing quantities and prices. If the original conditions were efficient, then taxes distort them and cause total output to shrink. Of course, the opposite may occur. The original conditions may have been inefficient, with prices not equaling true social values. If the tax just nicely corrects those original distortions (e.g., reaching Point *A* in Figure 5), the tax itself is *efficient.*

Incentives: How taxes may affect choices

Economists have long studied how taxes change people's individual choices. They have focused most closely on the *incentives to work.* Taxes may discourage people from working, by taxing away the money they are paid for their work. It is a perennial issue, and, indeed, it seems obvious to many people that taxes discourage work by cutting the rewards for it. Improving the economic incentives for effort has been a main theme of the Reagan economic program. But economic analysis shows that

the ***incentive effects*** are not obvious at all. Taxes may encourage work, discourage work, or not affect it at all.

The actual effect depends on each person's supply curve of labor, a curve that we discussed in the chapter on labor. Two contrasting cases of supply curves are shown in Figure 6, for Ludwig and Jewett. The amounts each will work per day are related to the wages they can obtain. Imagine that both Ludwig and Jewett are paid $8 per hour and that they both happen to work an average of 7 hours per day.

Now suppose that a 25 percent tax on income is imposed, so that their net pay after tax is only $6 per hour. That after-tax income is shown by the horizontal line at $6, which cuts the labor supply curves at Point *A* for Ludwig and *B* for Jewett. Evidently, the tax leads Ludwig to work less but Jewett to work more. The general conclusion is easy to grasp: The effect of an income tax upon the incentive to work depends on the shape of the labor supply curve. Depending on the curve, the effect may be positive, negative, or neutral.

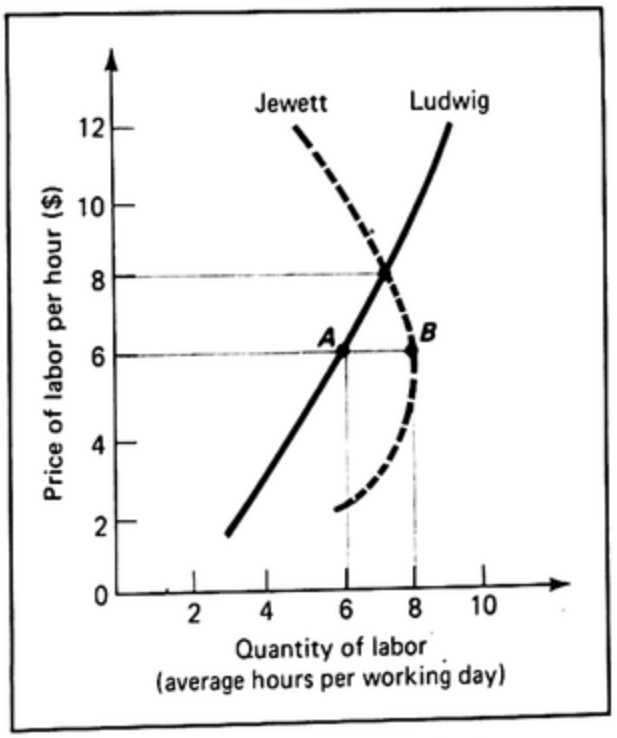


Figure 6 Analyzing the effects of a tax on incentives to work

At a wage rate of $8 per hour, both workers choose to work for 7.0 hours per day, on the average. A 25 percent tax would reduce their take-home pay to $6 per hour. In response to that tax, Ludwig cuts back to 6 hours per day, while Jewett increases to 8 hours per day.

If the original wage rate were $4 per hour, the 25 percent tax would lead both workers to reduce their hours of effort. But if Jewett's original wage rate is $10 per hour, even a 50 percent tax would cause him to increase his hours.

Many students find it easier to understand Ludwig's response than Jewett's. Ludwig works less because he is paid less. But Jewett's response can be explained by recalling the *price and income effects* from Chapters 6 and 14. As the net wage rate falls, it makes work less attractive at the margin (the price effect). But it also cuts the income that the worker receives for a given amount of work (the income effect). If the person needs the income desperately enough, the income effect will overcome the price effect, causing that person to work more. Jewett is such a case for wage rates over $5, as shown by the bending back of his supply curve above the $5 level.

But notice that Jewett's curve slopes upward at wage rates below $5 per hour. Therefore, if the original wage rate is $4 and the 25 percent tax cuts the net wage to $3 per hour, both Jewett and Ludwig will cut their work times. This makes the whole effect of taxes on people even less predictable because the same person may respond differently at differing wage levels.

Altogether the work-incentive effects of taxes are an open issue, even if the tax rates are high. Thus, even a 50 percent rate would lead Jewett to work more, if the original wage rate were $10 per hour. The same logic also applies to any other kind of effort that taxes might discourage, such as inventing, investing, or the creative arts. It also holds for estate taxes, where the parents' efforts to provide wealth for their children might be cut by taxation.

Taxes may either discourage or encourage effort. The outcome is a matter of

degree. Despite its obvious importance to the whole work burdens and productivity of the populace, the issue has not been settled by factual research. Economists have not been able to determine which way the actual effects go, on balance. As in other areas, the logic is clear, but the magnitudes are difficult to unravel.

Tax friction Nevertheless, taxes do clearly alter the choices of many taxpayers. Even if some people work more, many others probably work less. Some people shift toward barter and more do-it-yourself projects. Other choices are also disturbed. For example, people buy more houses because the interest paid on mortgages is tax deductible. Business managers may spend more because they are on tax-deductible expense accounts. Every day millions of choices are shifted because of tax effects.

Economists call the resulting distortions of choices ***tax friction.*** This friction subtracts from national output because it disturbs the efficient conditions that tax-free market choices would have given. Many of the disturbances are small, but together they can add up to significant totals.

Economists have tried to devise "costless" or "frictionless" taxes that won't alter any choices. But the search has been discouraging. Virtually all taxes are keyed to the levels of what is taxed: income, sales, property values, and so on. The closest to an ideal frictionless tax is the so-called lump-sum tax, which is levied simply on a person or enterprise. It is carefully *not* based on any economic magnitude. Rather, it is just a dollar amount taken either once-for-all or at some interval. Thus, a tax of $500 per year on each adult person would be a lump-sum tax. It would not specifically alter people's choices about work. But, of course, even this lump-sum tax would cause friction by making people poorer and therefore leading them to change some of their choices.

Perhaps even better in principle is a tax on pure *rent.* When any good is in perfectly inelastic supply, then all payments to it are rent. A tax on pure rents to land would be nearly frictionless, for it would not affect the supply of land. To that extent, taxes on land values are superior to other taxes because they do not distort choices at the margin. But land taxes do have practical problems because most land has capital improvements, whose supplies *are* disturbed by property taxes.

Distribution: The effects of taxes and spending We now arrive at one of the most basic features of public finance: how it affects the distribution of income and wealth. At each moment, the population is arrayed along a spectrum between low and high economic status, between poor and rich. Each person's status is determined by such *economic* forces as the income from work and the inheritance of property from parents. The state also affects this status by its *policies,* in applying taxes that take money away and spending programs that provide money or valuable services.

In short, taxes and spending have an incidence on the whole distribution of income and wealth. This is a larger topic than the incidence of sales taxes between sellers and buyers, discussed above. Now the basic question is whether each tax or spending item moves income up or down.

Progressive and regressive incidence The crucial concepts are progressivity and regressivity. A ***progressive tax*** falls more heavily on the rich than the poor. A ***regressive tax*** is the opposite, taking more heavily from the poor than the rich. Spending programs follow the same logic. Economists also speak of the incidence of taxing and spending taken together: Does the

whole budget tend to shift money and benefits down the income scale or upward? Even if one specific tax is regressive, the net effect of all taxes and spending may be progressive.

The term *proportional* is the key to defining progressivity and regressivity. A proportional tax takes the same percentage share of income, no matter what the level of income is.* A constant 15 percent tax rate, for example, is proportional. It is illustrated in Figure 7 by Line *A*. The horizontal axis is the amount of money income. The vertical axis is the tax rate, as a percentage of income. The flat-rate 15 percent proportional tax is a horizontal line, for the percentage rate does not change as income changes. The *marginal* tax rate out of each additional dollar of income is identical to the *average* tax rate out of all income. Both are constant at 15 percent.

A progressive tax takes a rising share of income, as income rises. Line *B* in Figure 7 illustrates such progressivity, using recent actual rates of the U.S. federal income tax as they apply to a single person. The tax rate rises by steps: The first $2,300 of income has a zero rate; the next $1,100 pays a 14 percent rate; the next $1,000 a 16 percent rate. Then the rises get steeper; between $14,000 and $41,500, the rate rises from 26 percent to 49 percent. The rises continue on up to a maximum 70 percent rate for income above $108,300. These rising rates contrast with the constant rate of the proportional tax.

The rising tax share as income increases shows a "progressive incidence." Progressive incidence tends to reduce inequality, by making after-tax incomes more equal. Neutral incidence leaves inequality unchanged: Everyone's income is scaled down by the same degree.

The progressive stair-step rates in Figure 7 are called *marginal tax rates* because they show what share is taken out of each marginal increase of income. There is also an *average tax rate* at each level of income. With a proportional tax, the marginal and average rates are identical throughout, as at 15 percent in Figure 7. But the average rate diverges from the marginal rate when progressivity occurs. Below a $2,300 income, both rates are zero; but by $3,400, the marginal rate is 14 percent and it is pulling up the average rate. The average rate lags below the marginal rate, for it also includes the earlier zero rates on income below $2,300. The same principle holds throughout, because the average rate includes all earlier rates, while the marginal rate applies strictly to the last dollar.

Note that even when marginal rates are 50 percent, the average tax bite may still be substantially lower. For example, at a taxable income of $85,000, the marginal rate is 50 percent. But the tax actually paid would be $40,000, which is an average of 47 percent. In short, the high marginal rate may have a large bite, but the taxpayer's whole sacrifice may remain less sharp on average.

Now consider the case of *regressive taxes*. They take a lower percentage of income as income increases, as illustrated in Figure 8. Several major taxes are, in fact, regressive. Sales taxes, which are typically 4 to 8 percent of the sales dollar as levied by most states and some cities, fall on people's consumption expenditures. Those purchases are a bigger share of income for poor people than for the affluent. Thus, a New York family with $7,000 per year may have to spend it all on consumption items, paying the 8 percent sales tax on all its income. A $100,000 New York family might spend "only" $30,000, saving the rest. The

*Strictly speaking, the structure of a tax can be considered separately from its effect on actual after-tax incomes, that is, from its incidence. But economists commonly use the term *incidence* for both meanings.

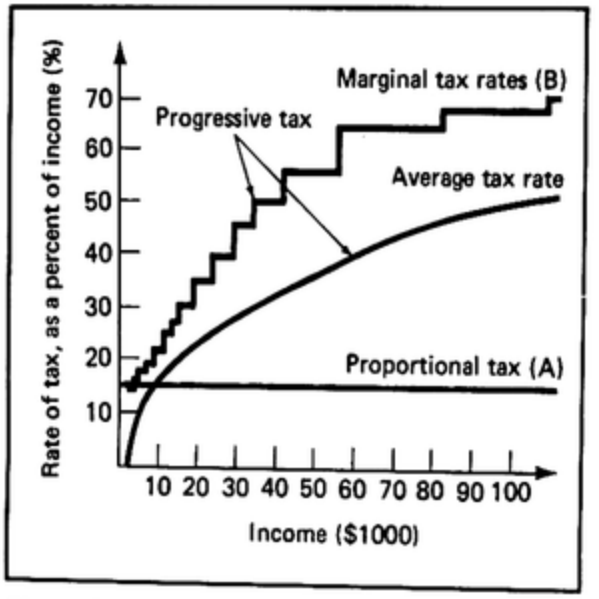


Figure 7 Proportional and progressive tax rates

The proportional tax is a constant percentage income, regardless of level. A proportional tax at a 15 percent rate is shown by the horizontal line.

A progressive tax imposes steeper rates as incomes rise. The marginal rates follow a rising stair-step pattern. The average rate of tax actually paid is below the marginal rate, as shown. The progressive tax rates shown are the recent U.S. federal income tax rates for a single taxpayer. The taxable income is the amount after all deductions.

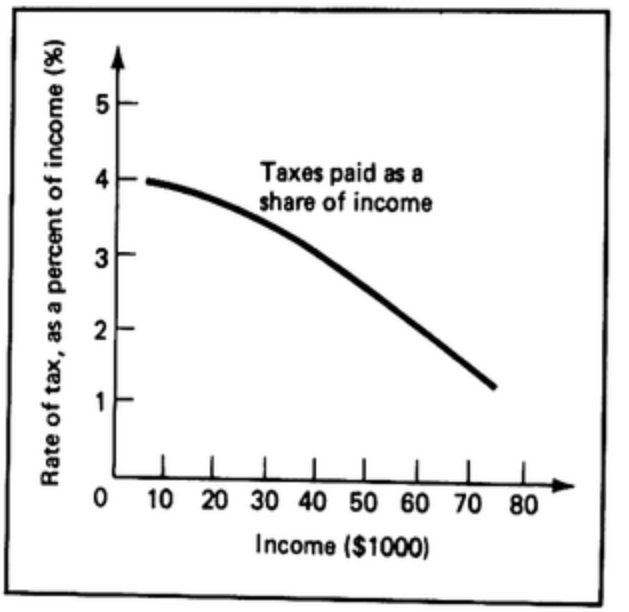


Figure 8 A regressive tax

The sales tax is levied at, say, 4 percent of sales. Families with low incomes spend virtually all of their income on consumption goods, so that the 4 percent tax takes away about 4 percent of their incomes. But more affluent families spend a smaller share of their income on retail items, so their sales tax payments are a smaller share of their incomes. A $60,000 family spends "only" $30,000; the 4 percent tax on that $30,000 yields $1,200 in taxes, which is only 2 percent of the $60,000 income.

resulting $2,400 that it pays in sales tax on its $30,000 of purchases is only 2.4 percent of its income, compared to the 8 percent paid by the $7,000 family.

Other regressive taxes include gasoline taxes, liquor taxes, and real estate taxes. Each falls *proportionally* more heavily on the poor than on the rich because, on average, these items loom larger in the budgets of lower-income people. (Note: Renters also pay real estate taxes, but only indirectly in their monthly rents. The owners pay the taxes directly, but pass at least some of the tax on in higher rents.)

Public spending programs also have an incidence on the distribution of income by providing money and other benefits unevenly along the income scale. Some programs are tied firmly to income tests, so that they go to poor families. Examples are income supplements for the poor, and public housing that is provided at below-cost rents to families with low incomes. Such programs are progressive because they tend to equalize the distribution of income. Other programs are regressive because their benefits go mainly to upper income people. Farm policies, for example, have channeled large payments to large-scale farmers, most of whom are already prosperous.

In sum, the basic logic of incidence is straightforward. Society can tax and spend in ways that move money and benefits up or down the income scale. Logically, too, economists can define an *optimum incidence* of these policies, reflecting any given set of social goals. But when one tries to

apply such broad judgments in practical details, there are two problems to solve:

1. *How much* progressive redistribution would be optimum is highly controversial.
2. The optimum policies will involve careful *balancing* among various kinds of taxes and spending.

The task is often complex because the goals are controversial. Moreover, the actual incidence of policies is often hard to discover because people manage to escape or twist the policies in unexpected ways. Also, the policies are not created by benevolent wizards; they evolve instead in an imperfect democratic political process.

Economists have defined the main features of a "good" tax system as follows:

1. The distribution of tax burdens is fair, by whatever criteria the society chooses.
2. Taxes are chosen and designed to minimize friction, which reduces the economy's efficiency.
3. The system is understandable to taxpayers and run at as low a cost as possible.

With these goals in mind, you can now approach the actual patterns of taxes and spending.

Major patterns of public finance

Among all the varieties of spending and taxes, there are several important trends and patterns. We present first the trend of total spending and then the composition of spending and taxes. In the process, we will also explain such technical features as indexing and tax expenditures. We save for a later chapter a review of how progressive the incidence of taxes and expenditure really is.

Economic policies apply rules and controls, as well as taxes and spending. Those rules and controls are too varied for a full survey, but some of them will be shown in the three case studies in Chapter 20.

Trend and share

The expanding economic role of governments is reflected in the trend of total U.S. public expenditure in Figure 9. A rise in the dollar totals was inevitable: As the population grew, national output increased, and prices rose. To put that dollar growth in perspective, Figure 9 shows total government spending as a share of gross national product (or GNP). That share was below 20 percent of GNP until 1940. It rose sharply after 1940 because of World War II and the large military expenditures due to the "cold war" of the 1950s. Then the 1960s brought major "Great Society" welfare programs, the space race to put people on the moon, and the Vietnam War. These ventures swelled spending, leading after 1980 to the Reagan administration's efforts to reduce the size of government. In short, there have been specific forces and conditions at work in the United States, causing both the rising trend and a series of changes. Yet, total government spending since 1953 has only risen from 28 to 33 percent of total economic activity. Less than one third of GNP now passes through the hands of governments in the United States. Actual government purchases of goods and services have remained at about 20 percent of GNP since 1952.

Is that "high" or "moderate"? By international standards (see Table 3), the United States is below the middle range of industrial economies in this respect. The relative affluence of many of these countries permits them to afford high levels of various public services. However, the sim-

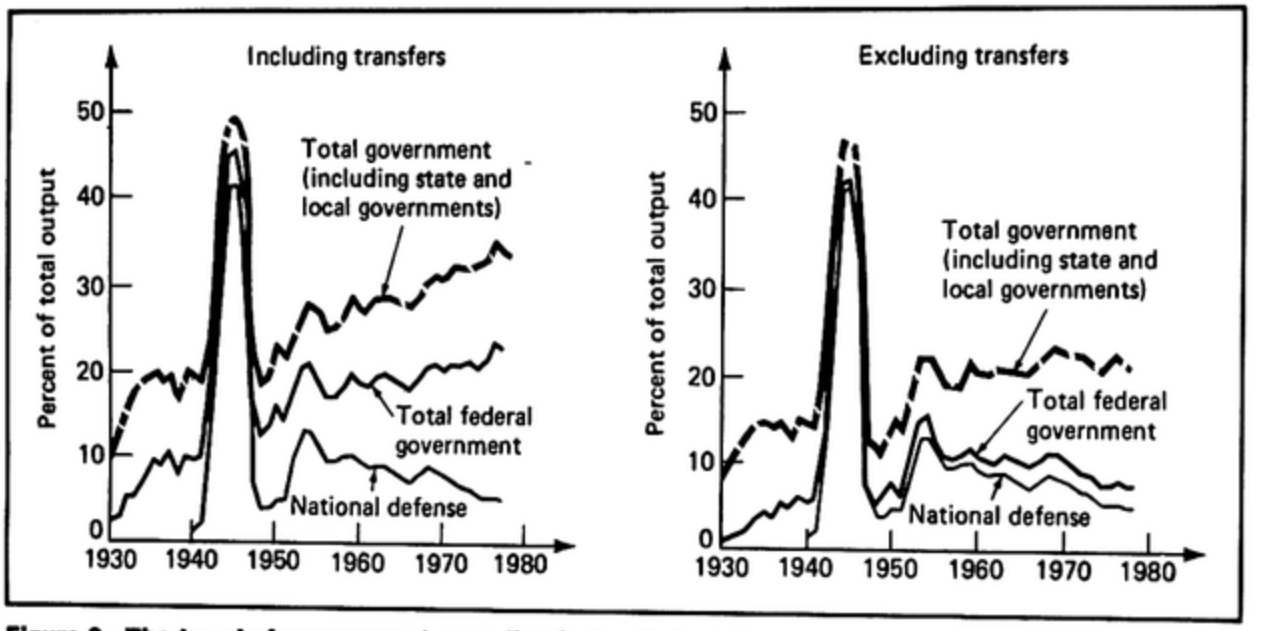


Figure 9 The trend of government spending in the United States

Before 1930, total public spending was below 10 percent of GNP. Programs to cure the Depression raised it toward 20 percent, and then World War II increased it sharply to nearly 50 percent. Since 1950, total spending has risen from about 25 to about 33 percent. Yet, public purchases of actual goods and services have stabilized at about 20 percent of GNP.

ple comparisons offered in Table 3 do not indicate whether in each country all the public programs are close to their efficient levels. Social needs differ sharply from country to country, because of differences in social structure, military needs, geographical size, and natural resources. Also, national preferences differ about what society should provide for its people.

Table 3 The relative size of the public sector in selected countries

Country	Tax Revenue at All Levels of Government, as a Percent of GNP (1980)
Sweden	56
Netherlands	55
France	45
United Kingdom	43
Canada	39
Switzerland	36
West Germany	30
United States	28
India	20
Japan	18

Sources: Tax revenue: International Monetary Fund, *Government Finance Statistics Yearbook*, Vol. V, 1981. GNP: International Monetary Fund, *International Financial Statistics*, December 1981.

Table 3 suggests little more than that the United States' share of GNP is lower than that taken by governments in other industrialized Western countries. The reductions in certain nonmilitary programs by the Reagan budget cutters in 1981–1984 will reduce the share slightly. In the 1940s, Colin Clark argued that any rise of public spending above 25 percent of GNP would cause severe economic problems. The experience of most western economies has belied that warning. Public expenditure can go too far and may be wasteful in many cases, but no general law seems to hold. One must look instead at the parts, to see if they are efficient and fair.

Composition: Local, state, and federal shares

The growth of spending by the three main levels of government is shown in Figure 10. Until the 1930s, spending by local governments was more than half of the total, while spending by the federal government was slight (except during World War I). The ruling conservative doctrine regarded federal programs as unnecessary, even reckless. All true needs could be met lo-

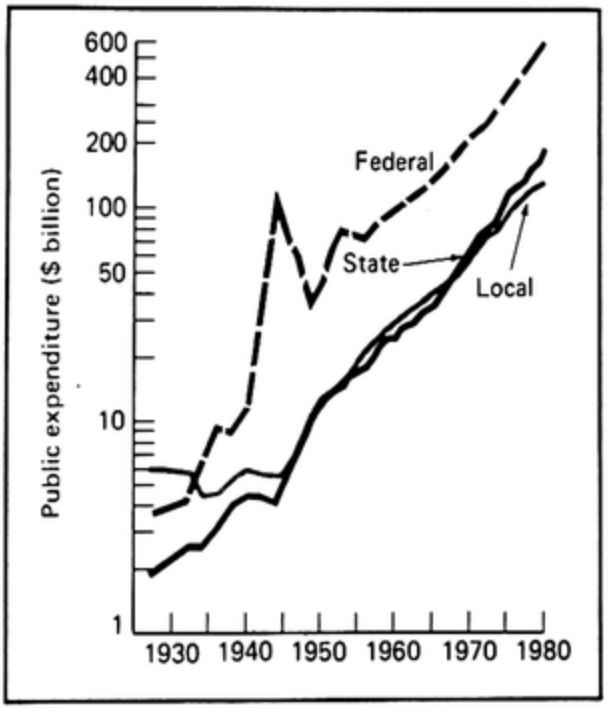


Figure 10 Public spending by the three U.S. levels of government

Local expenditure was originally the largest. Then federal spending rose sharply and has continued to grow. State spending has played an intermediate role.

cally, it held. Income taxes played a minor role, since they were held to invade personal freedoms.

Local spending has continued to be important throughout modern times. Its main purposes are basic ones: schools, streets, utilities, and personal safety. The tie between taxing and spending is close, for the funds stay in the locality. Yet, there has been a growth of local programs paid for mostly by the federal government, including aid to the poor, urban renewal, and health care.

State spending has traditionally been the least important, but since 1965, it, too, has expanded strongly. "Revenue sharing" has channeled back substantial amounts of federal funds to the states, which the states can use as they wish.

Altogether, since 1940, federal spending has been the dominant share, but local and state spending have been keeping pace since 1965. Thus, issues of public policy are important at all levels.

Purchases and transfer payments

Most public spending goes to buy work, products, and services. Such *purchases* from private markets cover an endless variety of items, such as workers' skills of various types, military weapons, school buildings, asphalt for roads, garbage trucks, and gasoline for police cars. The flow of these items now totals 20 percent of GNP, as shown in Figure 9.

Transfer payments have risen sharply since 1965, as Figure 11 shows. The rise reflects partly the growth of "Great Society" programs designed to relieve poverty and urban problems. Social Security payments and other pensions also rose after 1968, when Congress "indexed" them to the consumer price index. When price inflation occurs, these payments automatically rise.

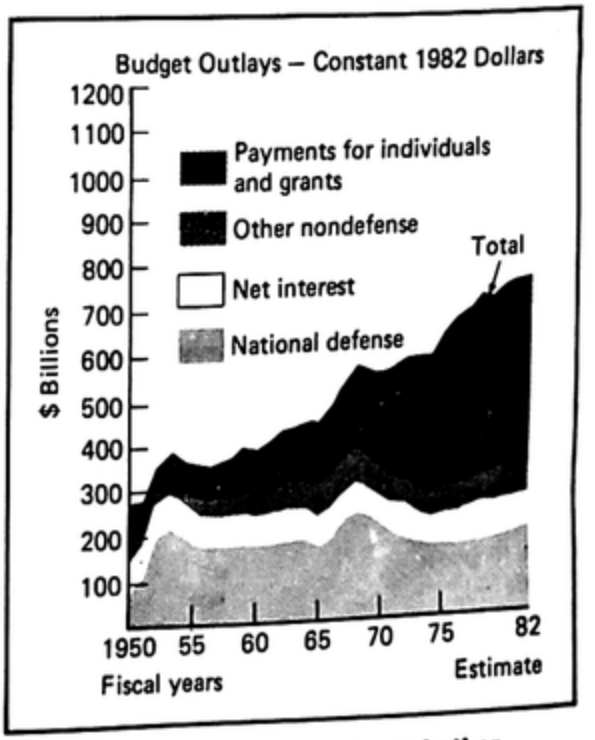


Figure 11 Transfer payments and other federal government outlays 1950–1981

The main transfer payments are in the upper part of the trends in the figure. Their share has grown since the early 1960s.

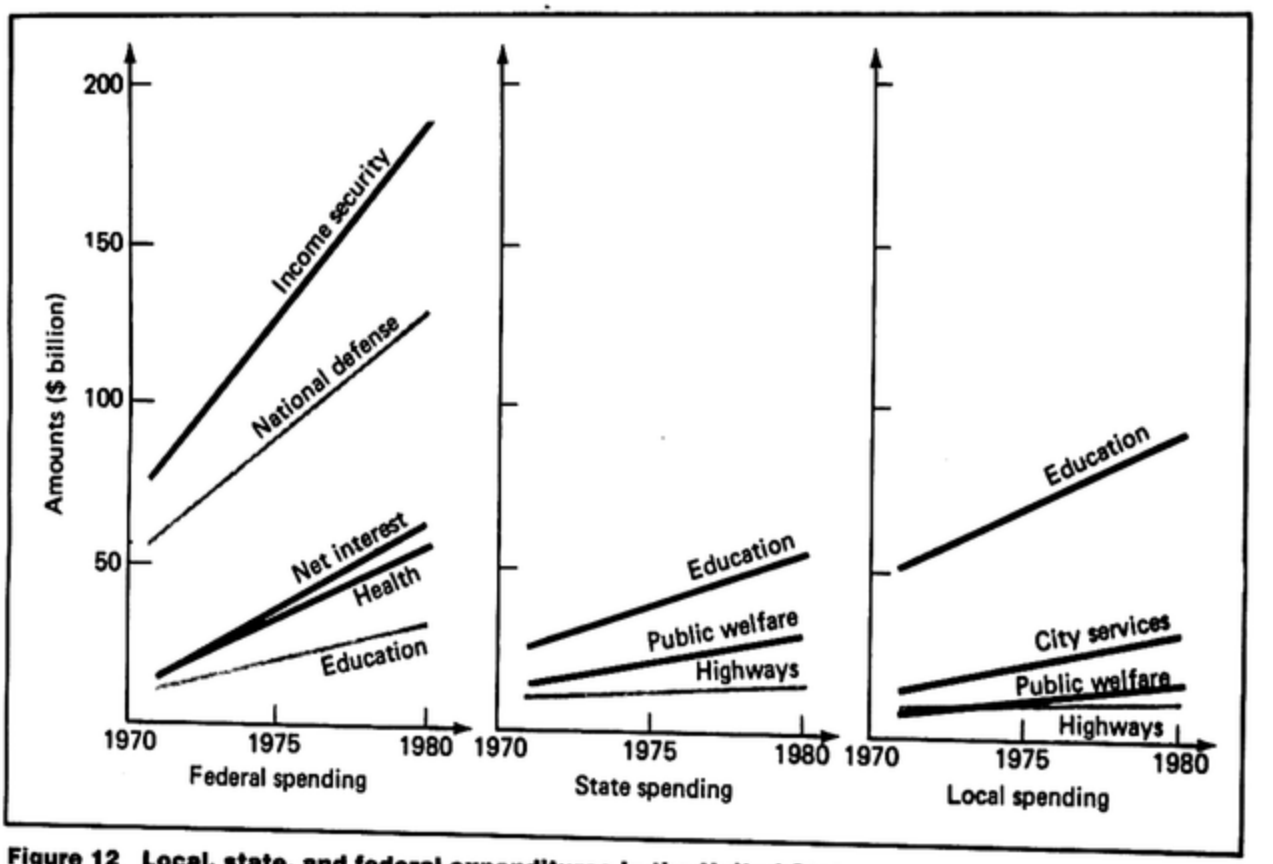


Figure 12 Local, state, and federal expenditures in the United States

However wise these growing transfer payments may be, they do not take goods out of the private economy—they only shift money to different consumers. Transfer payments, therefore, are less important than purchases in altering the capacity and composition of national production.

The variety of spending programs

Figure 12 indicates the diversity of the main spending programs at the three levels of government. Federal programs range from agriculture to space, military defense, highways, welfare payments, and Medicare. State spending is mainly for education and roads. Local spending goes primarily to schools, roads, and local services.

Several important programs are not shown in the budget. They are financed "off-budget," out of so-called trust funds with money from earmarked taxes. One such off-budget item is the federal highway program, which is paid out of gasoline taxes. Another is large loan-guarantee programs to farmers and to cities for pollution-control equipment. Their inclusion would appreciably alter the federal budget.

The loan guarantee programs need a special word. Perhaps they seem like a costless method for the government to induce people to make investments that it wants. Thus, to promote city development, the government guarantees loans to companies locating in the desired cities. The firms get the loans at lower interest rates, cities get the factories and jobs, and with luck, the loans are paid off. There is no direct public expenditure.

However, three defects mar this neat logic. One is economic: The loan guaran-

tees actually do have a significant opportunity cost. Each government must pay to borrow money at going interest rates (governments currently issue over $1,500 billion in bonds). The more it borrows or guarantees, the more the government draws funds away from private uses and from other worthy public purses. Since that raises the cost of funds, the government must pay higher interest rates. This increase is the cost of the "costless" guarantees. It can be large, even though it is indirect.

The other two defects are practical. Usually, the guarantee program spreads to cover a variety of loans unrelated to the purpose of the program. Thus, the Federal Farm Loan programs in the 1970s made loans for golf courses, condominiums, and shopping malls. Many other such programs become windfalls for interest groups. Finally, some of the loans don't change behavior even though they formally go to firms making the correct investments. If the worthy investment would be made even without the loan, as is often the case, then the loan guarantee is a waste.

Taxes

Turning to taxes, one also finds great variety and uncertain effects. The main lines are given in Table 4 for federal, state, and local taxes. Federal income and Social Security taxes have become dominant now at 60 percent of all taxes and 20 percent of total GNP. Next are the various sales taxes, and then property taxes. The corporation profits tax is highly valuable over the business cycle, supplying between 10 and 25 percent of federal revenue. The estate tax, which is intended to reduce the impact of inherited wealth, continues to represent only 2 percent of federal taxes and less than 1 percent of all taxes.

Tax expenditures ("loopholes") We need also to discuss ***tax expenditures.*** The standard tax rates do not, in fact, apply to everyone. There are many exceptions and special provisions, so that taxpayers'

Table 4 Taxes and revenues at the local, state, and federal levels (% billions)

	1932	1950	1970	1980	
				Amount	Percent
Local					
Property tax	4.2	7.0	33.0	64.1	29.9
Sales and gross receipts tax	0.26	0.48	3.1	9.3	4.3
Individual income tax	—	0.64	1.6	4.1	1.9
Other	1.7	8.0	51.4	137.0	63.9
State					
Sales and gross receipts tax	0.73	4.7	27.3	58.3	25.9
Individual income tax	0.07	0.72	9.2	29.1	12.9
From federal government	0.22	2.3	19.3	51.2	22.8
Other	1.5	6.2	33.1	86.4	38.4
Federal					
Individual income tax	0.41	15.7	90.4	181.0	42.0
Corporate income tax	0.60	10.5	32.8	60.0	13.9
Sales and gross receipts tax	0.73	7.8	18.3	25.5	5.9
Other	0.90	6.1	64.1	164.8	38.2

Sources: *Statistical Abstracts*, various years; *Historical Statistics*; *Census of Governments*.

Table 5 ***The main federal tax expenditures***

	Amounts (in billions)	
	Corporations	Individuals
Lower tax rate on capital gains	$ 1.0	$ 22.3
Deductibility of state/local taxes paid		17.7
Deductibility of medical and health costs		15.7
Investment tax credit	15.4	3.1
Exemption of contributions to pension funds		15.1
Deductibility of consumer and mortgage interest payments		14.9
Exemption of Social Security and similar benefits		14.2
Deductibility of charitable gifts	1.0	8.0
Lower corporate tax for smaller firms	6.9	
Exemption of state-local bond interest	4.7	3.0
Other	13.5	25.7
Total	$42.5	$139.7

Source: U.S. Office of Management and Budget, *Special Analysis of the U.S. Budget for 1980*. The totals shown are of dubious validity because of interactions among the loopholes shown, which would change the amounts of some if others were eliminated.

guidebooks to federal income taxes run over 100 densely printed pages. These loopholes reduce the standard rates for various groups. Some exceptions reflect a variety of valid social and economic purposes. But others merely reward affluent people who can exert political leverage. At any rate, these special provisions reduce the true progressivity of taxes.

The technical term for them is "tax expenditures." The phrase reflects that standard tax revenues are "expended" by reducing what certain groups would have to pay. Thus, each taxpayer receives a $1,500 deduction from taxable income per child. The tax expenditure for this deduction, multiplied by the 60 million children in the United States, deducts $90 billion from income. Since the average tax rate on that income would have been about 20 percent, some $18 billion in potential taxes is left in certain taxpayers' pockets.

Certain local and state bonds receive another large tax expenditure, because the interest they pay is exempt from federal taxes. This helps cities to issue bonds for local schools, roads, hospitals, and the like, since they do not have to pay the going market rates of interest. However, it has grown into a large windfall for affluent people. Though anyone can buy any of the billions of such tax–exempt bonds issued annually, only those people in marginal tax brackets over 45 percent gain from doing so. Below an income of about $40,000, the after-tax yield from a taxable bond is better than that from a tax-exempt bond. But the comparison of after-tax yields reverses at higher income levels, and thus at higher marginal tax rates. Accordingly, high-income people can put most of their assets in tax-exempt bonds and avoid nearly all federal income taxes. On balance, the benefit to cities must be compared to the regressive effects on distribution.

Tax expenditures are indirect, and their amounts can often only be estimated. The main tax expenditures are shown in Table 5. All are "loopholes," but all are

also legal and many have valuable effects. Economists insist that they involve real costs, much as if the public money were actually spent directly. The $180 billion of tax expenditures in 1980 needs to be included in any evaluation of public finance.

SUMMARY

The policy issues in public finance are often complicated, but economists have developed several basic concepts and analyses to clarify them.

1. A social good is nonexclusive in consumption. The total demand for such a good is a vertical summation of the individual demands. This total demand interacts with supply to determine the efficient amount of the public good.
2. A public interest arises when economic decisions create external effects. The effects can be costs or benefits.
3. Other economic reasons for public policies include common-property resources, monopoly, and unfair distributions of wealth, income, and opportunity.
4. Under efficient policies, the marginal benefits per dollar spent on alternative public programs and on private goods will be equal.
5. Cost-benefit analysis is a method for defining the efficient amount of specific public programs. It equates marginal social costs and benefits. It does not tell who should pay for the program: the beneficiaries or the general taxpaying public.
6. Spending divides into purchases and transfer payments. Taxes are of personal and *in rem* categories.
7. The incidence of an *in rem* tax usually depends on the relative elasticities of demand and supply.
8. Taxes may affect incentives to work, to invent, or to do other activities. The direction of effect depends on the shape of the individual supply curves of effort.
9. Tax friction is the loss of production that occurs when people alter their decisions so as to lighten their tax burdens. Such friction is the real economic cost of taxation.
10. A progressive tax takes a rising share of income as the taxed variable increases. Therefore, the tax falls more heavily on the rich than on the poor. A regressive tax is the opposite, taking disproportionally more from the poor.
11. Actual taxes and spending take about one third of GNP in the United States. They make up a complex patchwork at the local, state, and federal levels. Tax expenditures are amounts that would have been collected in taxes if special exceptions had not been made.

Key concepts

Efficiency, equity, stabilization
Social good
External effects: costs and benefits
Public expenditure
Cost-benefit analysis
Transfer payments
Taxes
Incidence
Incentive effects
Tax friction
Progressive and regressive taxes
Tax expenditures

Questions for review

1. a. Give five examples of a social good.
   b. Why are private markets likely to supply few social goods?
   c. Ideally, how would a government determine how much of a public good should be provided to its citizens? What are the two chief drawbacks to this ideal process? Explain.
2. a. What is the rule that will lead to maximum net social benefit from each public project? Explain.
   b. Why must future benefits and costs be discounted when calculating the costs and benefits of a project?
   c. Should the private sector interest rate or the rate at which the government can borrow funds be used to determine the appropriate discount rate? Explain.
3. a. Define: regressive tax, progressive tax, proportional tax. Give an example of each.
   b. If a country's tax system is regressive, does its budgetary policy shift money from the poor to the rich? Explain.

19

Inequality, Poverty, and Discrimination

As you read and study this chapter, you will learn:

- why income and wealth are unequally distributed in the United States
- why some people are poor despite our high standard of living
- how discrimination affects economic inequality
- how taxation, government spending, equal opportunity laws, and the minimum wage affect the distribution of income

At its best, the market economy rewards skill, hard work, and creativity. However, it also penalizes people who are unable or unwilling to behave in economically productive ways. Most Americans seem to accept this as fair in some sense. At least it is free from corruption.

But basing incomes on marginal productivity does not produce equality. Labor skills are unequally distributed, and the ownership of land and capital is even more unequal. Thus, even if income were strictly based on productivity, it would be unequally distributed. Discrimination only reinforces a tendency already built into the market economy. Deliberate government policies alleviate inequality, but they do not eliminate it.

This chapter is divided into three main parts. The first describes the income and wealth distribution in this country and outlines some of the reasons for inequality. The second discusses the impact of discrimination based on race and gender. The third

analyzes various government policies that are meant to make the income distribution more equal.

Income differences and their causes

In this section, we look at economic inequality and its basic causes.

The degree of inequality

The basic patterns are shown in Table 1. The U.S. population is arranged from the lowest to the highest levels of wealth (what people own) and income (what they earn). One can compare the shares of income and wealth held by the poorest one fifth of the families with those of the richest one fifth. The same families, roughly speaking, will be found at the same positions on the wealth and income scales. Wealth provides access to income, and the two are closely related.

As Table 1 shows, wealth is much more unequally distributed than income. That has long been true. A relatively few families can accumulate much wealth, but most scrape by with little or no net assets. In recent years, the richest 1 percent of families have held about 25 percent of all private assets. Taking just corporate securities (stocks and bonds), the top-ranking 1 percent of the population held 57 percent of the value of all personally held securities in 1972. The richest 5 percent of the populace held nearly 70 percent. The top 100,000 families held over $3 million in assets each. Their average wealth was $15 million per family. Such wealth permits families to combine a very high level of consumption with continued growth in wealth, while work is a matter of choice. The very rich can live in a state of affluence that the ordinary citizen cannot easily imagine. The richest 5 percent of the populace are largely free from financial anxiety. Their lives differ sharply from the lives of those at the bottom of the income and wealth scale.

At the lowest end of the income distribution are the approximately 10 percent of families that live in poverty. They receive only about 2 percent of total income, and they hold less than none of the assets because they are in debt. The average income of the top 10 percent of families is about 15 times as large as the average income of

Table 1 The distributions of income and wealth in the United States

Population groups	Income share 1978	Average income, $ 1975	Wealth share* 1962	Wealth share* 1972	Average wealth, $ 1962	Average wealth, $ 1978
Bottom fifth	5.2	5,222	0.2		169	
2nd fifth	11.6	11,649	2.1		1,776	
3rd fifth	17.5	17,574	6.2		5,245	
4th fifth	24.1	24,203	15.5		13,112	
Highest fifth	41.5	41,677	76.0		64,289	
Highest 5%	15.6	62,666	—		—	438,230
Highest 1%	—	—		24.1		693,942
Highest ½%	—	—		18.9		

*Wealth is defined as net worth (assets minus debts).
Source: *U.S. Statistical Abstract*, 1980; and U.S. Congress, House Committee on the Budget, *Data on the Distribution of Wealth in the United States*, September 26, 29, 1977.

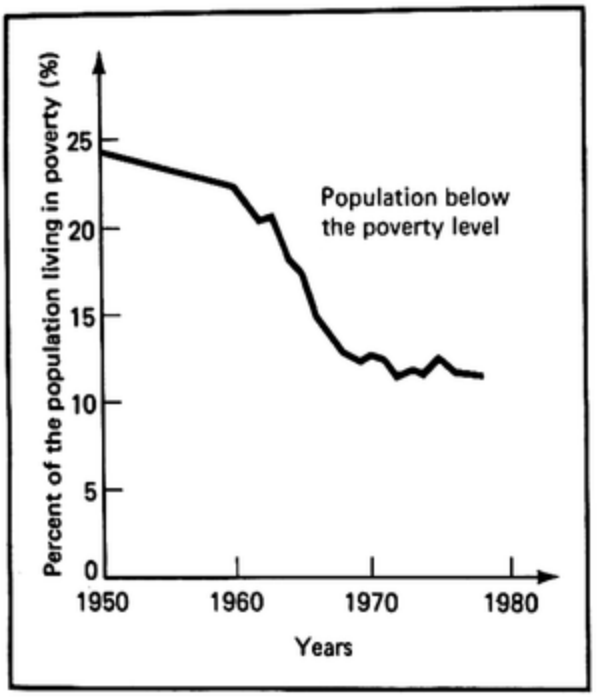


Figure 1 The level and trend of poverty in the United States since 1950

About one in ten American families lives in poverty, a total of about 25 million people. The percentage of Americans living in poverty is slowly declining, although the 1970s brought little change.

Source: U.S. Statistical Abstract, 1980, p. 464.

the bottom 10 percent. Having no net assets, the average poor family is subject to financial insecurity from job loss, sickness, and accident.

Trends In the 20th century, there has been a trend toward a more equal distribution of wealth and income but it has been slow. The extremes of income narrowed slightly from 1920 to 1950, but have scarcely changed since then. The inequality of wealth narrowed distinctly during 1930–1949, but there has been little change since 1949.

Mobility Within the distributions of wealth and income, there is some mobility. In particular, there is a surprising degree of *upward* mobility. **The dominant characteristic of great wealth in the United States is that it is usually created rapidly.** By a spectacular success, the founder of a dynasty builds up immense wealth in a decade or two. One of our cultural myths is that wealth is accumulated slowly, by saving from income over a lifetime or two. But, in fact, most great wealth has come into being quickly, created by innovations, discoveries, monopolies, patents, and luck.

Once it is created, wealth tends to persist. Approximately 50 percent of the largest fortunes derive from inheritance, not from new wealth. This persistence of wealth provides stability rather than change. Such families as the Rockefellers, du Ponts, and Mellons have been wealthy for generations. Families with long-established wealth are at the top of the social structures in New York, Philadelphia, Boston, and most other large cities, except in the oil-rich Southwest.

At the lower economic levels, poor families are caught in a continuing cycle of poverty, poor education, and inferior jobs. Although children often move up the ladder, and some make large leaps, most people who are born poor stay poor.

Poverty Although "poverty" has no fixed definition, a series of thorough government studies has led to a general agreement on an income standard below which people are clearly living in poverty. By that measure, poverty continues to be a major problem in the United States.

Figure 1 shows the extent and trends of poverty in the United States over the past 30 years. There is a long-term downward trend, though there was little change in the 1970s. About 25 million Americans, just over 10 percent of the population, still live in poverty. Such poverty translates into a hard, grinding life, often marked by family strain and hopelessness. Coexisting as it does with the high affluence of part of

the populace, such poverty is very divisive socially.

The poor are not a single homogeneous category. There are several distinct groupings of poor people.

Over 30 percent of *black families* have incomes that fall beneath the poverty line. They suffer from high unemployment rates, especially among the young. Over half of poor blacks live in families headed by a woman. *The elderly* account for nearly one fifth of all those living in poverty. They often have inadequate pensions, or none at all. Though many are able and willing to work, they cannot find jobs. *Single-parent families* are a special category, accounting for about half of all poor people. The strain on the parent in such cases is severe. She or he must struggle to earn a living and to manage the household. The children often lack adequate emotional support and guidance, which hinders their own development. *Farmers* are a heterogeneous lot, but about one sixth of them are poor, especially those on small farms in the South. Other categories of poverty include people living in economically depressed regions, people with low intelligence and skills, and immigrants who face language problems and discrimination. Altogether the poor constitute a diverse group with widely varying characteristics, although these often overlap in specific individuals. Thus, a black family on a small southern farm, with grandparents living in but no father at home, is especially likely to be poor.

Poverty is a major problem that tends to be transmitted across generations. It is partially alleviated by special public programs such as ADC (Aid to Dependent Children) grants, food stamps, and Medicaid and Medicare. After allowing for the benefits of such programs, only about 6 percent of families remain below the poverty line. Yet, nearly every city and large town has a sizable group of poor people; New York alone has over a million. Because of the diversity of its causes, poverty is unlikely to yield to any single cure.

Technical causes of apparent inequality

Economists have long known that some of the apparent inequality in the distribution of total wealth and income stems from purely technical aspects of how it is measured. These sources of inequality need to be filtered out, to arrive at the true degree of economic inequality.

Age and life cycles Earnings usually rise with age, in the life-cycle pattern of increasing wages. Thus, in 1978, families whose head was below 28 years of age had incomes averaging $12,500, while those with heads aged 45 to 54 averaged twice as much ($25,400). A 20-year-old clothing salesman may make only $10,000 now, but he might expect to make $30,000 by the time he is 45. A junior executive makes $18,000 and has few assets at age 25. But her 50-year-old counterpart makes $70,000 and has $200,000 in assets. The younger person has a good chance to reach the older colleague's level. To assess true inequality in income and wealth, one should compare people *of the same age.*

Regions Because of differences in climate alone, regions differ in costs of living. The warmer climate in the South and Southwest and on the West Coast makes housing and clothing costs lower. Such factors can make a 30 percent difference in living costs. Also, cities are generally more costly to live in than towns or farms. Among cities, too, there are sizable differences in costs. Manhattan is a much costlier place to live than Muncie, Indiana.

Family size The adequacy of a family income is affected by the size of the family. For example, if there are just two adults,

an income of $15,000 will be much more adequate than if there are seven children plus a grandparent in the family.

The economic forces shaping the income distribution

Some factors that cause inequality improve economic efficiency, some are neutral, and others cause inefficiency.

Causes improving efficiency Talent and effort are the two main personal characteristics that affect equality. Both influence the distributions of income and wealth, but neither is the dominant cause of inequality. Effort is almost neutral in its effect on the inequality of wealth. Many people in low-wage jobs work long, hard hours, but their intensive effort often produces income barely above the poverty level. Immigrants are the classic example of unusually hard workers. Yet, their strong efforts commonly elevate them only from the lowest levels of income to the lower middle classes. Effort alone, then, does not explain much of the total disparity in wealth and income.

Creative talent is probably more important than effort in explaining inequality. Many large, rapidly created fortunes have come from major innovations, such as instant photography by Edwin Land and xerography by Chester Carlson. These people captured some of the extra value created by their genius (and efforts), but some of the benefits were also passed on to consumers. Thousands of lesser fortunes have arisen from patented inventions and from the creative building of businesses. Yet, such purely creative activity has probably not been the major source of wealth. Many of the most creative people, including inventors, have worked for a salary for firms or public agencies. Their contributions have not made them personal fortunes; instead, the benefits have gone mainly to others.

Luck Luck operates capriciously through people's genes, location, timing, and other factors beyond personal control. Examples of "lucky" wealth are stock market winnings and successful commodity speculation. Much oil industry wealth is also a matter of luck. On the negative side, personal accidents, floods, droughts, and speculative disasters frequently separate people from their assets and income. Even if one could magically create perfect equality today, differences in luck would restore a large degree of inequality tomorrow.

Causes reducing efficiency People have found many ways to exploit their fellow humans. Economists recognize several methods by which economic power has enriched some people at the expense of others: A good example is monopoly power.

Monopoly not only reduces economic efficiency, it also impairs equality. It shifts income and wealth from the average citizen to a few people who hold unusual amounts of capital. One fairly reliable estimate of the monopoly effect on distribution suggests that 20 to 40 percent of the private wealth held by the top 5 percent of families probably came from monopoly power. Many large family fortunes can be traced back to the exercise of market power in some particular industry. Such instances include the Rockefellers (oil), the du Ponts (chemicals), and the Mellons (oil, aluminum). Some of the wealth acquired in this way was later given to universities, which were named for the "robber barons" whose money built them. Examples include Stanford University (Leland Stanford—western railroads), Duke University (James B. Duke—the American tobacco monopoly), Vanderbilt University (Cornelius Vanderbilt—railroads), Carnegie-Mellon University (Andrew Carnegie—steel; Andrew Mellon—oil and banking), and Rockefeller University (John D. Rockefeller—oil).

Discrimination

Discrimination involves the differential treatment of people on the basis of superficial, easily perceived differences. Despite the ugly connotations of the word, not all discrimination is undesirable. For example, employers routinely discriminate among job candidates on the basis of their past experience and recommendations from former employers; readers select books according to their expectations of the contents.

Discrimination is unfair or undesirable if the perceived differences do not in some sense objectively justify the differing treatment. For example, race, sex, and ethnic background in themselves have no systematic effects on worker quality. Therefore, discrimination based upon such criteria is economically harmful.

Employment discrimination

When women, blacks, Hispanics, native Americans, and other groups are subject to job discrimination, it undermines equal opportunity. The discrimination has two elements—exclusion from employment, and low pay. First, people are *excluded from certain jobs:* They are not permitted to apply, or are not hired when qualified, or are hired only in trivial numbers. Second, even if hired, they are *paid less for equal work,* relegated to inferior status, and not promoted in a timely manner.

In practice, employment discrimination of both types has been common in many markets. The earnings differential between men and women, as illustrated in Figure 2, reflects both the exclusion of women from a wide range of jobs (a trend that did not begin to decline until the 1960s) and sharp differences in pay rates. Until the 1970s, women were simply kept out of most heavy work, skilled craft jobs, and management positions. Women's "place" was said to be that of clerk, telephone operator, salesgirl, grade school teacher, and cleaning lady. Blacks are still confined largely to the menial, unpleasant jobs. So are Hispanics and native Americans. The pay for these groups is 30–60 percent below that of white males.

Discrimination has eased somewhat since the 1960s, more than had seemed possible before then. Yet, the total gains have been small. Job inequalities not based on skill or experience still exist, and minorities and women are often only a token presence in many of the upper-level jobs (such as managers, financiers, doctors, and lawyers).

Part of the inequality is caused by inadequate skills, which many minority people have because of poor schooling, the ab-

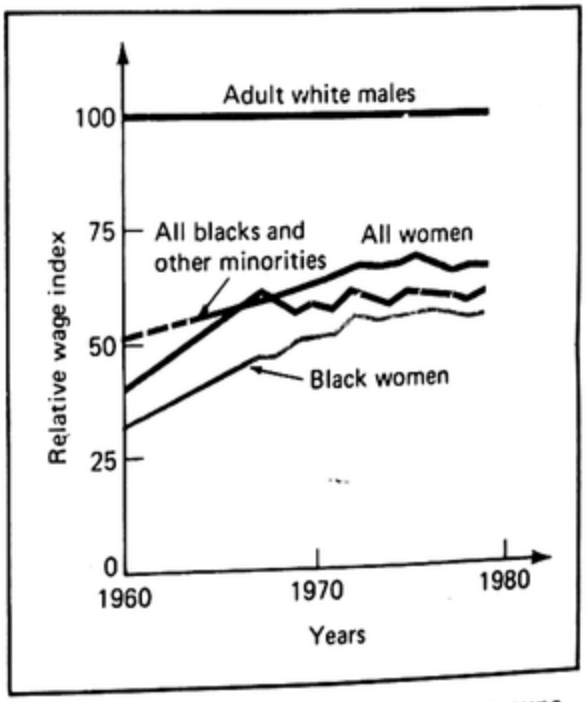


Figure 2 Relative earnings of various groups within the U.S. population 1960–1978

Earnings by adult white men are the standard of comparison; they are set at an index value of 100. Women and blacks have been paid at just about 60 percent of the income levels of white men. Black women are paid even less. The gaps narrowed moderately during the 1960s but remained steady during the 1970s.

Source: Data from *U.S. Statistical Abstract,* 1980, p. 424, calculations based on average weekly earnings.

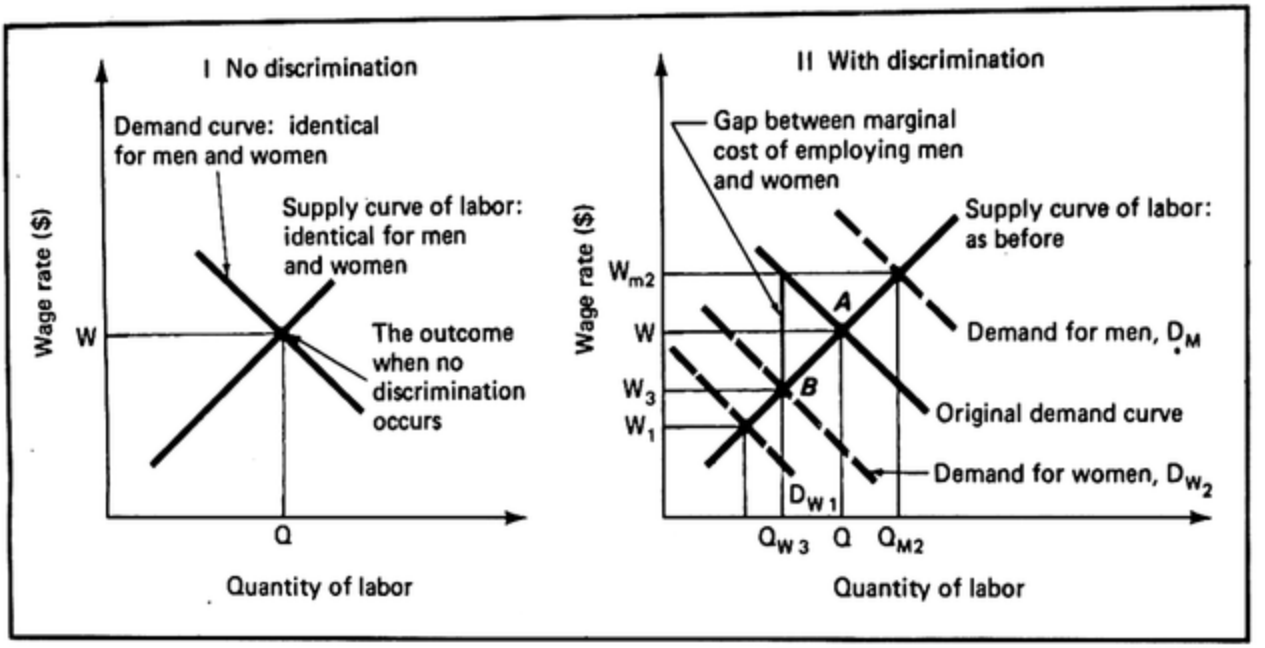


Figure 3 The effects of job discrimination on wages and employment levels

When opportunity is equal (Panel I), supply and demand curves are identical for women and men. They are hired in roughly equal numbers at the same wage rate.

Discrimination (Panel II) shifts in the demand curve for women, causing them to endure lower wages and a smaller level of employment. Men are now favored with more jobs at higher pay. The marginal cost of male labor is now W_2; of female labor, W_3. The gap between W_2 and W_3 is large, as shown. It reflects the inefficiency, as well as the unfairness, caused by discrimination.

sence of past incentives to gain skills, and the lack of adequate family financial support. The problem of inadequate skills will take time and resources to correct, although its importance is often less than is claimed. Many skills can be learned quickly on the job, if there are adequate incentives. Although exclusion from employment and lower pay are conceptually distinct parts of job discrimination, the problem can be usefully analyzed without separating the two effects.

The analysis of discrimination begins with simple supply and demand curves for labor in a typical job market. But now there are two groups of workers, women and men, seeking work as, say, construction supervisors. Assume that the women and men are equally qualified (that is, their marginal revenue products are identical), and that their supply curves are also the same, as shown in Panel I of Figure 3. Therefore, with no discrimination, they would be hired in identical numbers and paid the same wages (Point *A* in Panel II). The selection process between women and men would be random and sex-blind.

But if women are regarded as "unsuitable," the demand for female labor shifts in. If the decline is severe, as with the labor demand curve D_{W_1} in Panel II, then fewer women will be hired, and only at a dramatically lower wage, W_1. This often happens when women are deemed to be "out of place" in risky or unpleasant jobs, or in complex "responsible" jobs like running an investment bank, piloting an airline jet or double-bottom oil tanker, or presiding over a government agency.

A lesser degree of discrimination would move the labor demand curve for women in moderately, perhaps to D_{W_2} in Panel II of Figure 3. In that case, the results at Point *B* are less severe for the

women, but in equilibrium, women laborers receive less pay, and fewer are hired.

The demand for male workers shifted out when the demand for women shifted in, moving to Demand Curve D_M. This gives a new equilibrium for male labor, with more jobs and a higher rate of pay. To draw the extra men away from other jobs, where their marginal revenue product would be higher, the firms in this industry must bid up their wages. The new equilibrium is at a higher number of male workers, Q_{M_2}, and a higher wage, W_{M_2}.

Accordingly, men take jobs from women *and* the men's wages rise while the women's wages fall. Such a shift would cause sharp protests if it were to happen suddenly. In practice, however, it usually occurs gradually and by tradition, so that it is less visible and controversial. Sometimes, indeed, the effect gradually dwindles as time passes. But lost jobs and lower pay do occur. Exclusion can be complete or virtually so; and pay rates may differ a lot.

Discrimination is both inequitable and inefficient. This can be seen in Figure 3 by comparing men's wages with the cost of hiring more women. Men are being paid the wage W_{M_2}. Yet, the marginal man could be replaced by an equally productive woman at a cost of W_{M_3}, which is only about half of W_{M_2}. The marginal costs of labor are thus out of line with the marginal products. This distortion causes the discriminatory economy to produce less than would be possible with equal opportunity.

This disparity of wages might be expected to trigger corrective actions by competitive firms: They would hire the women, reduce their own costs, and gain extra profits. Such a process does occur in some degree, but it has not eliminated the main patterns of discrimination. When discriminatory attitudes are ingrained in the culture, they can resist the inducements of the profit motive.

Firms that exclude women and minorities often say that the excluded workers are simply not qualified. Thus, if all executives are white males, they explain that they could not find qualified women and blacks for these high posts, and that it would take decades to locate and train new candidates. For lower-level jobs on the production line, the minority applicants are often said to be unreliable and disobedient.

If that were true, the lower demand curves D_{W_1} or D_{W_2} in Figure 3 would reflect the lower marginal revenue product of such groups, and would not be the result of unfair discrimination against them. The firm might regret the results, but it would be discriminating on relevant economic grounds. In practice, some minority workers *are* technically less productive because of poor schooling, work habits, and so on. The key questions are: Is the quality difference truly present? How severe is it? How easily can it be overcome? Often the distinction is illusory or quickly remediable by on-the-job training.

Discrimination in housing

Minority groups have long been excluded from affluent neighborhoods and confined to ghetto areas of the major cities. Such discrimination in housing is sometimes subtle, but usually effective. Although it has declined since the 1950s, it is still widespread. The main economic elements are straightforward, and supply-demand analysis can clarify them.

Consider a typical neighborhood housing market without racial discrimination, as illustrated in Figure 4. The equilibrium price and quantity of housing exchanged are determined by the intersection of total demand for and supply of housing. The resulting quantities are divided so that Q_W houses are sold to whites and Q_B to blacks, on the basis of relative demand.

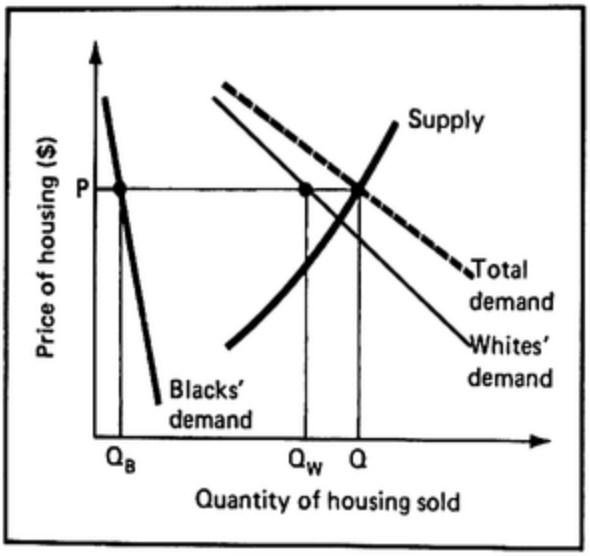


Figure 4 Housing sales without discrimination

With no discrimination in the housing market, Q houses are sold at a price of P. Q_B houses go to blacks and Q_W go to whites. Sales reflect buying power regardless of race.

Now suppose that racial discrimination against blacks occurs. That results in two separate supply curves for housing, contingent on the buyer's race. The supply to blacks is represented by Curve S_{B_1} or S_{B_2} in Figure 5; the supply curve to whites is S_W.

If exclusion is complete, then supply to blacks is S_{B_1} and the quantity exchanged is zero at all prices. Such a supply curve coincides with the vertical axis. Even at the highest price they would be willing to pay (at Point *A*), blacks cannot buy any housing in this neighborhood. Meanwhile, whites can buy housing at Point *B*. That gives a lower equilibrium price compared to the no-discrimination case because the discrimination has ruled out blacks as buyers. That lower price is endured by the *sellers* of the houses to keep blacks out. It is a price the white neighborhood will pay.

However, the temptation to sell at a price as high as *A* may finally stir some neighbors to sell to blacks after all. Some limited supply for black buyers may emerge, as in S_{B_2} in Figure 5. A few houses will be sold at Price *C*, which is above the price to white buyers at *B*. Once the rigid color line is broken in this way, market forces may drive the two prices together, and discrimination will be squeezed out as Prices *B* and *C* converge.

Discrimination is both unfair and economically inefficient. The effects are not always recognized because such practices are traditional and ingrained.

Public policy and income distribution

Poverty is deeply rooted and surprisingly resistant to political solutions, but four main types of public policy have been di-

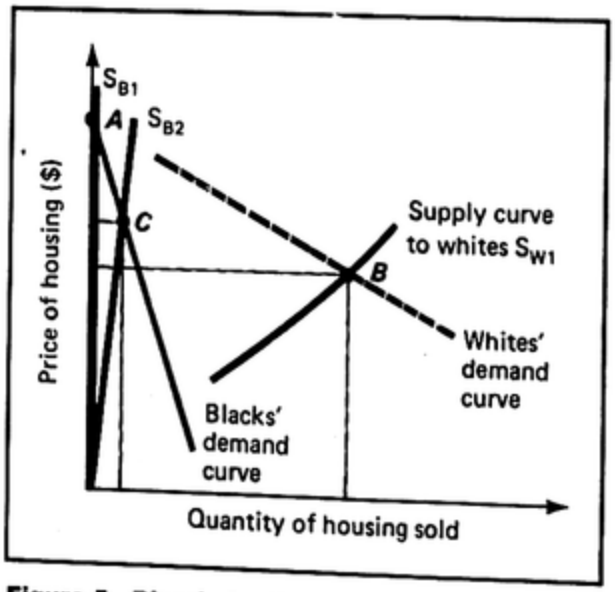


Figure 5 Discrimination changes the amounts and prices of houses sold

Discrimination shrinks the supply of housing to blacks, perhaps all the way to zero, as shown by S_{B_1}. Owners will not sell even if blacks offer a high price, at Point *A*. The supply to whites is now the same as total supply was before, so that the price whites have to pay goes down to *B*.

If some supply is offered to blacks, as shown by S_{B_2}, the price *C* results, where supply and demand for blacks are in equilibrium. Blacks can buy, but at a higher price than whites. Owners may then supply more houses to blacks, to get the higher price. This will tend to close the gap between *C* and *B*.

rected at changing the income distribution: progressive taxation, transfer programs, equal opportunity programs, and minimum wage laws. The first two have been discussed generally in Chapter 18. Only their effects on inequality are examined here. The last two policies are introduced and evaluated in this section.

Tax policies

Recall that *progressive taxes* take a higher *percentage* of income as income increases. Taxes could significantly even out the income distribution if they were sufficiently progressive.

A *regressive* tax takes a *lower* percentage bite as income increases. It accentuates existing inequality. Several major taxes do appear to be regressive, chiefly property, sales, and excise taxes. The regressivity occurs because the poor tend to spend a higher *percentage* of their incomes on housing, cigarettes, and other taxed consumption goods.

Finally, *proportional* taxes take a *constant* percentage of income or wealth and do not affect inequality.

The burden of taxes is ultimately borne by individuals. It is customary to define that tax incidence in terms of *disposable income**, which is what remains after taxes are netted out of gross income. There are three main types of taxes. Income taxes are the largest ($214 billion in 1980), but sales taxes ($93 billion) and property taxes ($64 billion) are also substantial.

Each household's disposable income is simply:

$$\text{Disposable income} = \text{Gross income} - \left(\text{Income taxes} + \text{Sales taxes} + \text{Property taxes} + \text{Other taxes}\right).$$

*This is a somewhat different concept of disposable income than that used in national income accounting.

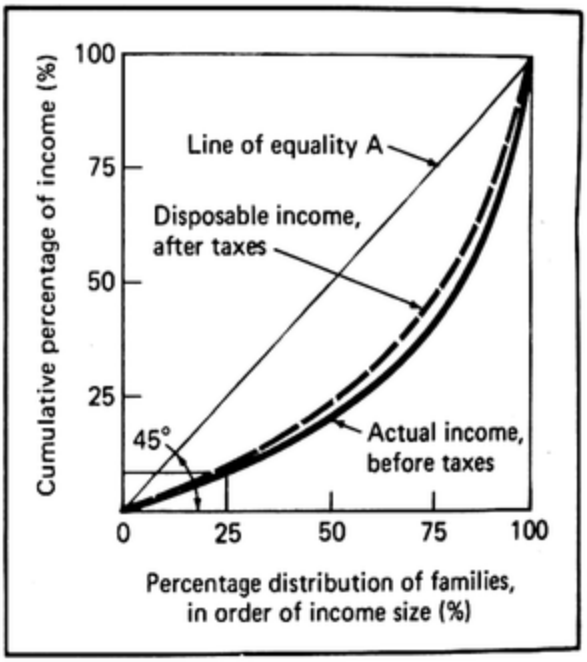


Figure 6 A Lorenz curve showing the degree of inequality

With households arranged in order of their incomes, a perfectly equal distribution of income would give Curve *A*. The "lowest" 25 percent of families would have 25 percent of the income, and so on. Actual pretax incomes are unequal, as shown by Curve *B*. The lowest 25 percent of families have only 10 percent of total income, etc. Taxes tend to equalize incomes slightly, as illustrated by Curve *C*, which is above Curve *B*.

The specific tax bite on each household will vary with many factors: income levels, amounts of goods bought, taxable property owned, and the tax rates being applied. The tracing of tax burdens is a highly complex matter, depending on elasticities and patterns of property ownership that are not accurately known. Income and property taxes are straightforward and their effects are fairly well known; for sales taxes, the estimates are less precise; and for other taxes, the effects can only be estimated.

If they were known perfectly, a "Lorenz curve" could be drawn, like that in Figure 6. The population is arranged along the horizontal axis in order of increasing income. The vertical axis is the cumulative share of their incomes, starting with the lowest income groups. If incomes were

perfectly equal, the distribution would be the straight line *A*. Any inequality will cause the actual distribution to lie along a curve below *A*, such as *B*.

Now suppose that Curve *B* is the actual distribution of income before taxes. Each household in that distribution is at one point along the horizontal axis. Then taxes are deducted, leaving disposable income. The resulting distribution of that disposable income is Curve *C*, again with each household located somewhere along the horizontal axis. If the net burden of taxes is progressive, then Curve *C* lies above Curve *B*, reflecting a more equal distribution of disposable income than of gross income. If taxes are instead regressive, then the disposable income curve would lie below Curve *B*.

Many studies of tax incidence in the United States and other countries have been done in recent decades. They involve varying assumptions about the shifting and incidence of various indirect taxes, and their results are not uniform. Table 2 and Figure 7 present the results of one major study, showing patterns that have probably changed little since the study was undertaken.

As expected, *federal taxes* are broadly progressive, as shown by the rising line (*B*) in Figure 7 and line 6 in Table 2. The regressive effects of the federal excise taxes on cigarettes and gasoline (line 4) and of the payroll tax (line 5) are outweighed by the progressivity of the income tax (line 1). State and local taxes are regressive, as shown by the down-sloping line (*A*) in Fig-

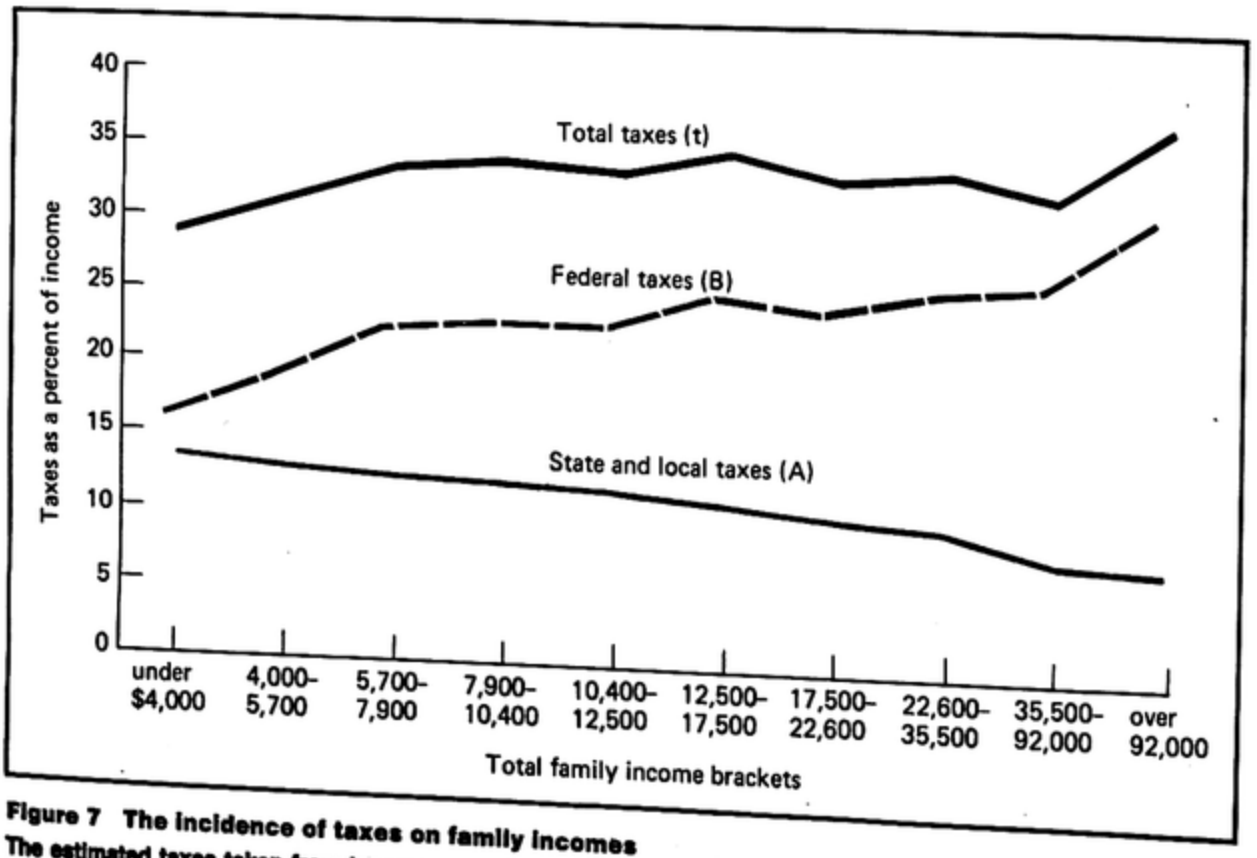


Figure 7 The incidence of taxes on family incomes

The estimated taxes taken from incomes are shown (as percentage shares) on the vertical axis. State and local taxes are regressive on the whole, as shown by Line *A*. Federal taxes are moderately progressive (Line *B*). Total tax incidence appears to be mildly progressive, especially at the lower and upper ends of the income scale.

Source: Musgrave and Musgrave, *Public Finance in Theory and Practice*, p. 267.

Table 2 ***Taxes as percentage of total family income, 1968***

Taxes	Income Brackets										
	Under $4,000	$4,000–$5,700	$5,700–$7,900	$7,900–$10,400	$10,400–$12,500	$12,500–$17,500	$17,500–$22,600	$22,600–$35,500	$35,500–$92,000	$92,000 and over	All Brackets
Federal Taxes											
1. Individual income tax	2.0	2.8	5.9	7.1	7.9	10.1	10.6	12.7	14.8	18.5	9.9
2. Estate and gift tax	—	—	—	—	—	—	—	0.6	2.0	2.7	0.4
3. Corporation income tax	5.1	6.1	5.0	4.0	4.3	4.6	4.8	5.1	5.3	6.6	5.0
4. Excises and customs	2.5	2.8	3.1	3.0	2.9	2.7	2.1	1.1	0.9	0.6	2.3
5. Payroll tax	5.5	6.3	7.0	6.9	6.7	6.1	5.2	4.2	1.5	0.6	5.2
6. Total	15.2	17.9	20.8	21.6	21.6	23.4	22.6	23.8	24.5	29.1	22.7
State and Local Taxes											
7. Individual income tax	—	0.1	0.3	0.6	0.7	1.1	1.4	2.3	1.6	1.3	1.0
8. Inheritance tax	—	—	—	—	—	—	—	0.2	0.6	0.8	0.1
9. Corporation income tax	0.4	0.5	0.4	0.4	0.3	0.4	0.4	0.4	0.4	0.5	0.4
10. General excise tax	3.4	2.8	2.5	2.3	2.2	2.0	1.7	1.0	0.5	0.3	1.8
11. Excises	2.7	3.0	3.3	3.0	2.9	2.5	1.9	1.0	0.8	0.6	2.1
12. Property tax	6.7	5.7	4.7	4.3	4.0	3.7	3.3	3.0	2.9	3.3	3.9
13. Payroll tax	0.2	0.5	0.8	1.0	1.0	1.0	1.1	1.2	0.2	0.1	0.8
14. Total	13.4	12.5	11.9	11.6	11.1	10.6	9.7	9.1	7.1	6.9	10.3
All Levels											
15. Total	28.5	30.5	32.8	33.1	32.8	33.9	32.4	32.9	31.6	35.9	33.0

NOTE: Items may not add to totals because of rounding.

SOURCE: Adapted from Richard B. Musgrave and Peggy Musgrave, *Public Finance in Theory and Practice*, 3rd ed. (New York: McGraw-Hill, 1980), pp. 267 and 275.

ure 7 and by line 14 in Table 2. Excises (lines 10 and 11), which are sales taxes, are highly regressive, as expected. Property taxes (line 12) are also regressive; they largely reflect taxes on houses, which are the only substantial assets that most lower-income families own. For higher-income families, housing is proportionally less important, and so property taxes take a smaller share of their income.

The combined effect of all taxes appears to be nearly proportional over the wide middle range of incomes, where over 70 percent of households are located. Toward both ends of the scale there is a higher degree of progression.

Public expenditures

Government spending, particularly transfer payments, can also have a significant redistributive effect. Recall that such expenditures can be considered progressive, regressive, or proportional, depending on which groups benefit most as a percentage of their incomes. For example, such programs as welfare payments, public housing, Medicaid, and food stamps are clearly progressive. Other expenditures are less progressive.

The value of many *specific* benefit programs can be assigned to income groups with a high degree of confidence. Transfer payments are the largest item, at a level of 7 percent of total income, followed closely by public education. General benefit programs are not so readily assigned, because their benefits are widely spread. Figure 8 presents one careful estimate of the incidence of benefits. The benefits appear to be much more progressive than the burdens (taxes), especially at the low end of the income scale, where transfer payments are highly concentrated. (The effects of these upon work incentives are considered in the adjacent box.)

The combined incidence of taxes and benefits ("Net" in Figure 8) is also highly progressive at lower incomes, thanks almost entirely to benefits. At about $8,000 income (in 1968), there is a crossover from net benefits to net burden. The poorer families receive net benefits; the richer families bear a net burden. Restated in current incomes, the crossover might be at about $20,000 income. Above that level, the progressivity is slight.

In short, public programs do markedly add to the real incomes of the poor as a group, but the degree of progressivity in other income ranges is mild. The poor are helped, but there is only a moderate tendency toward leveling down the affluence of the rich.

Equal opportunity programs

The 19th century brought economic opportunity to millions of immigrants in the United States, and legal freedom for the slaves. But women's opportunities were reduced on the whole, and those of native Americans were nearly obliterated. After the 1870s, segregation severely limited the opportunities of most blacks.

During the first half of the 20th century, economic discrimination against these groups remained strong. Only in the 1960s did government policies begin strongly to promote equal opportunity. The 1964 Equal Opportunity Act made job discrimination on the basis of sex or race illegal and created the Equal Employment Opportunity Commission (EEOC) to enforce fair hiring practices. After several years of experimenting, the EEOC centered its actions on large firms, to get a maximum economic effect for the least number of cases.

The EEOC usually reached compromises with discriminatory companies, requiring payments of money to minority

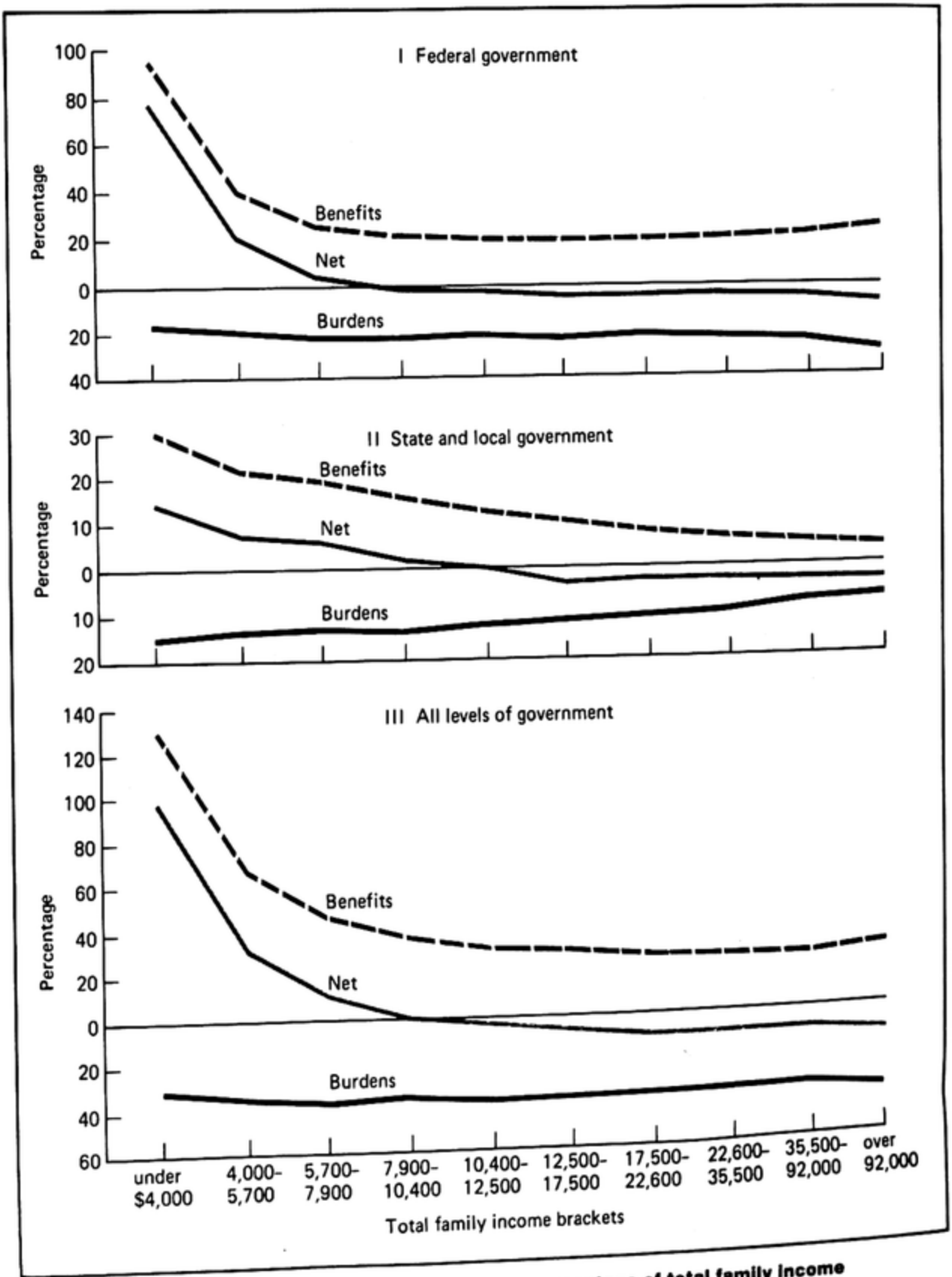


Figure 8 Tax burdens and expenditure benefits as a percentage of total family income

Source: Musgrave and Musgrave, *Public Finance in Theory and Practice*, p. 275.

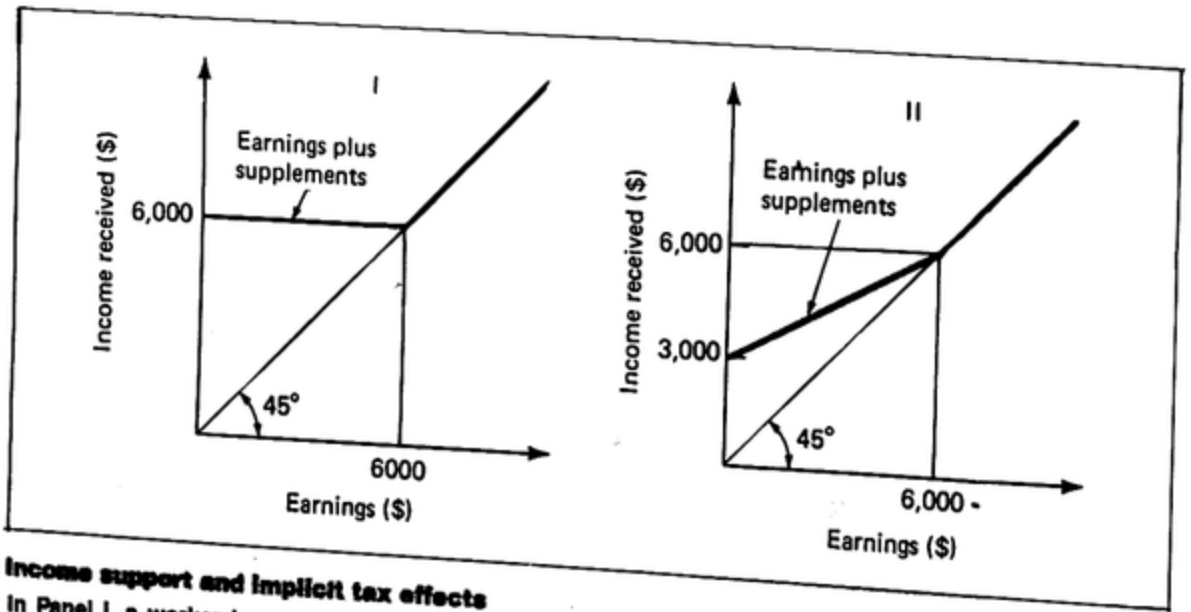
Welfare Payments and Incentives to Work

Few redistributive policies have raised more difficult economic issues than income maintenance programs for the poor (popularly called welfare). They include cash grants for Aid to Families with Dependent Children, as well as Medicaid, food stamps, subsidized housing and other benefit programs. Though they have grown rapidly since 1960, they have been common for many decades both in the U.S. and elsewhere.

The incentive problem they raise is fundamental. Because they are aimed at helping the poor, the benefits are confined to people below certain income levels. Any earnings which raise the family's income above that income ceiling are under a steep implicit tax, because they result in the loss of large benefits. Hence welfare may produce disincentives to work.

One version of the problem is illustrated in the figure. Earnings from work are measured on the horizontal axis; the total income received (earnings plus welfare benefits) are on the vertical axis. Earnings, therefore, trace out a 45° line, because they are earned and kept.

If $6,000 is a minimal income, society may supplement lower earnings so as to bring them up to that level. Panel I illustrates such a supplement program. The worker receives at least $6,000 even if earnings are zero. But notice that the supplement implicitly taxes below $6,000 at a 100 percent rate: The $6,000 is received even if no



Income support and implicit tax effects

In Panel I, a worker is guaranteed at least $6,000, regardless of his or her wages. Implicitly, income is taxed at a 100 percent rate up to the $6,000 level. In Panel II, the worker receives in transfers one-half the difference between $6,000 and his or her wages, up to an income level of $6,000. The implicit tax rate is only 50 percent.

work is done. This disincentive may have strong effects for workers whose lack of skills limits them to jobs in that range.

One alternative is to lower the income threshold, say to $3,000. Then the disincentive no longer operates in the $3,000–$6,000 range. But people genuinely unable to obtain work will now undergo hardship because they will receive only $3,000.

Alternatively, the supplement can be only partial, providing 50 percent of the difference between earnings and $6,000, as in Panel II. Then the implicit tax is only 50 percent. But many families will still fall below $6,000, and there is also a disincentive from the 50 percent implicit tax rate.

This approach is often called a "negative income tax." The progressive tax rate on higher incomes is simply extended below a threshold income level, to provide partial subsidies to poor people.

All welfare programs face the same central dilemma: *The disincentive effect varies directly with the degree of assistance provided.* As long as aid is based on an income test, there is an implicit tax. The greater the aid, the steeper the implicit tax, with its possible disincentives.

The effects of disincentives are not precisely known, but research has suggested that primary earners (e.g., the head of the household) are not sensitive to them. Secondary earners (e.g., working children) are more responsive, enough to cause significant declines in total earnings because of the disincentives.

and female employees to compensate them for having been underpaid in the past, and new programs to increase minority and female employment. The payments were only a crude way to offset the previous low pay. They missed all potential employees who had *not* been hired. Yet, the whole effort did substantially improve opportunity, especially for women. Meanwhile, individual cases piled up by the thousands, entailing long delays. Here the gains that the EEOC could obtain were small and specific, rather than large and broad. On the whole, the EEOC allocated its small resources fairly effectively, trying to get the largest total gain in opportunity.

Minimum wage laws

We presented an analysis of the effects of minimum wage laws in Chapter 5. They raise wages for some people but leave others unable to get the jobs they would prefer. The specific effects depend mainly on the elasticities of demand and supply of labor. Here we merely summarize the probable effects as of 1982. The Reagan administration has proposed sharp changes in the law, which could alter the effects.

The law's coverage has large gaps. Exemptions include household workers, trainees, farm workers, students, outside sales workers, all workers in retail and service shops with annual sales below $325,000, and handicapped workers. The law, therefore, does not affect about 10 million nonsupervisory workers. Many of those exempted are in the classic low-wage jobs, such as migrant farm workers and teenagers. Moreover, millions of other legally covered workers are, not, in fact, affected because their employers ignore the law.

Thousands of immigrants (both legal and illegal) toil in sweatshops at subminimum pay, unable to request legal wages and afraid to report their employers for fear of losing their jobs.

Altogether, 3 million workers are paid at the minimum wage level. About 3.9 million receive less, under legal exceptions. And probably 3 million more are paid less than the minimum illegally. The best estimates are that only about 5 million workers' wages are actually raised by the law. The increase probably averages about 35 cents per hour, which is about $700 per year. Perhaps half a million workers are forced to take other jobs because their marginal product is below the minimum wage rate. Another 100,000 workers may be unable to find work at all because of the wage floor. Those negative effects of the law fall most heavily on teenagers and unskilled workers:

> . . . teenagers have more to lose than to gain from higher minimum wages; they appear to be forced out of the better jobs, denied full-time work, and paid lower hourly wage rates. . . . If one of the goals of minimum-wage legislation is to eliminate sweatshop low-wage jobs, for teenagers the law appears to be counter-productive.*

Adult women, by contrast, "are the main beneficiaries of increases in the minimum wage. . . . A higher minimum brings adult females from the part-time into the full-time labor force, forcing even lower-wage teenagers out into the part-time jobs that they have vacated."†

Yet, high unemployment rates among youths, especially black males, may not be caused by the minimum wage, though one frequently hears arguments to that effect. Instead, many of these youths live in poor urban areas, where jobs simply do not exist at any wage rate.

All told, the law probably improves incomes by about 10–12 percent for 3 million low-wage workers; it reduces wages marginally for about half a million other workers; and it may price about 100,000 workers out of jobs altogether. Therefore, the effects of the law have been mixed.

*Edward M. Gramlich, "Impact of Minimum Wages on Other Wages, Employment and Family Incomes," Brookings Papers on Economic Activity, No. 2, 1976, Washington D.C. p. 409.
†Ibid., p. 462.

Summary

Equity in distribution is a complicated subject, with many unexpected, sophisticated issues that ultimately cannot be wholly resolved. These main points should be remembered:

1. Actual distributions of wealth and income in the United States are markedly unequal, though less so than in most comparable countries. Wealth is much more unequal than income, and much of the inequality is transmitted from generation to generation. The degree of inequality has been declining gradually.
2. Poverty is focused especially among minorities, single-parent families, small farmers, and old people. Inheritance is a substantial element in inequality.
3. Some inequality has been caused by differences in effort and creativity. But large fortunes arise primarily from other "instant" causes, including luck.
4. Discrimination has strongly affected the economic position of women and blacks and other minorities. It reduces efficiency as well as equity.

5. Government taxes and spending can have a progressive, neutral, or regressive effect.
6. Actual taxes are moderately progressive, with the basic progressivity in the federal income tax nearly offset by regressivity in sales, gasoline, liquor, cigarette, real estate, and other taxes. Government spending is more clearly progressive, mainly because of programs to aid the poor (welfare, food stamps, Medicare and Medicaid, etc.), which have grown since 1960.
7. Policies to equalize opportunity have had some effect. Minimum wage laws have probably helped lower-income workers on the whole, especially women. But teenagers and certain other poor groups have been hurt by them.

Key concepts

Discrimination
Benefit programs

Questions for review

1. Compare the economic inequality of two families: The Smith's income is $15,000 per year; the Jones's income is $50,000 per year. What factors would have to be taken into account before you can make an accurate comparison?
2. Use supply and demand analysis to explain how job discrimination against minorities will affect the wages and the numbers who are hired of both minority and non-minority workers in a given labor market.
3 Explain why discrimination in hiring that is not related to differences in productivity is economically inefficient.
4. The Federal income tax is progressive. Can it be assumed, therefore, that the net effect of taxes in the U.S. is a greater equality in income distribution than would otherwise be the case? Explain.
5. How do government expenditures affect the equality of income?

20

Education, Social Regulation, and the Military

As you read and study this chapter, you will learn:

- how educational policies pose issues of efficiency and equity
- how simple cost-benefit analysis can clarify the effects of policies to control pollution and industrial hazards
- how military policies—purchasing, arms levels, and the draft—can be improved along economic guidelines

Consider how far you have come in microeconomics. You first mastered the analysis of demand and supply. Next came costs, market outcomes, and supply in competitive markets. The effects of monopoly followed, and then the main inputs, capital, labor, and natural resources. Finally, you learned the main lines of public finance and studied poverty and discrimination. Now you are prepared to apply this training to three difficult economic problems facing society. These problems are the education of the young, the protection of safe living and working conditions, and national defense.

These three complicated cases can be clarified using basic economics. Indeed, the cases are a test of your skill at this analysis. They also test your maturity and sense of balance, for there are no simple answers. It is rarely clear what the strictly best policies are, for they often must strike a balance among several

goals and interest groups. Economists often show the effects of each policy, so that society can make informed choices.

Education, our first case, has been your own main job for the last 12 to 15 years. It is crucial both for each student and for the whole structure of society. We will show the main private and social elements that education contains and the questions of fairness that it poses. The first section of this chapter may lead you to see the economics of your college activity in new ways.

The second section discusses the protection of the environment from pollution. The great environmental cleanup in the United States since 1965 has applied several kinds of rules and incentives. Some of them have been effective, others not. We will show their nature and effects.

Then we take you through the remarkable economics of military spending. The Department of Defense spends over $160 billion per year. We show how the process departs partly from the conventional conditions of market efficiency, so that a degree of inefficiency is likely to occur. We also use simple theory to show why the arms race continues. It appears probable that there is too little competition among the armaments suppliers and too much between the United States and Soviet Russia! Finally we turn to the military draft showing why a volunteer army is usually more efficient.

In all three cases, the problems involve both the concepts of competition and monopoly and of public finance. Some of the social elements arise because monopoly conditions have strong effects. Indeed, the public programs themselves are often monopolistic in ways that limit people's choices and stifle their incentives for effort. Therefore, one must analyze monopoly and incentives both in these cases and in the budgets themselves. The *design* of the policies is as important as their *size*.

Economics of education

Educating the young has always been a leading social task. Some societies mainly teach religion and obedience to their children. Others try to instill creativity and independent thinking. All industrial societies now absorb most of their youths' time from the age of five to the mid- or late teens in primary and secondary schooling. Most young adults then take paying jobs, but a minority go on—as you have—to college, and a few more take advanced studies.

Although education differs sharply from country to country, it has the same economic features everywhere. It adds skills, which make people more productive. It sorts people into work that fits their skills, along the lines of comparative advantage (recall the brief discussion in Chapter 2). It instills traditions and methods, so that citizens can be stable members of society. Some teachers also do creative research, which helps to improve technology and enrich the culture. Altogether, education creates large economic values, ranging from technical productivity in factories and offices to the progress of knowledge itself. What you may have regarded as merely schoolwork is, in fact, part of a crucial economic process.

Education's various values divide into two classes: private and social. Each student gains *private* benefits by learning how to think more maturely and to do certain specific job-related tasks. Later, these benefits can be translated into higher pay on the job. Since the additional amounts of pay go only to the worker, they are strictly private. The *public* benefits are often more subtle.

We will now analyze both categories.

Private benefits of education

If there were no formal schooling, children would still develop productive skills as

they matured. But effective schooling raises the levels of those skills and thus provides higher productivity.* That, in turn, usually results in higher incomes for the educated person. These incomes are received and enjoyed privately. In exchange for spending all those years in school, you will eventually attain higher pay during your working life. Moreover, your jobs will probably be more pleasant, and you may understand the meaning of life more fully and be better able to cope with the complexities of modern existence.

Public benefits of education

There are three main types of *public benefits* which people's education may provide to other people rather than just to themselves. **First, universal public schooling provides everybody with the *basic skills*** for work and self-care. It enlarges the pool of productive workers and provides a variety of skills and mobility in labor markets. Because workers are more effective and intelligent, industry becomes more efficient and profitable. It can provide goods to consumers at lower cost. The whole society benefits, not just those who invested in an education.

Without schooling, many people would be unable to cope with the complexities of society. Some might turn to crime and violence; others would simply require public assistance. Indeed, many do both of these things now, and better schooling might well reduce their numbers. Society might relieve the economic and social burdens of crime by supplying better education to these citizens.

Second, schooling provides a *stable electorate*, reasonably well informed and able to function in an effective democratic process. Because people understand public issues better, they will deal with them more intelligently. Extremists are less dangerous to the fabric of a well-educated, skeptical society. The sense of decency and cohesion is deeper, and the ability to resist specious political claims is stronger.

Third, the greater productivity of the population provides a larger economic base for the *taxes* that finance *adequate public services.* Although education yields higher *private* incomes for people, these incomes are the economic base for meeting social needs. If productivity were lower, society would have fewer resources for all of the programs covered in Chapter 18, including education. Therefore, partly as a circular matter, education enlarges the scope for efficient public spending in the future.

These main public benefits are hard to measure precisely because they are so complex. Much of their value in forming the basic productivity and stability of the populace is provided in grades 1 to 12. That justifies subsidizing schools at those levels with public funds, perhaps with a total public subsidy.

For education at the college level, however, the public benefits may be relatively weaker when compared to the large, private career rewards. True, there is much valuable research done at universities, which gives public benefits that justify public funding. Moreover, the career system generates many of the future leaders in politics and society. Also, the taxes on the careerists' later earnings may have especially high yields for the public purse.

Nonetheless, the balance between private and public benefits in higher education is tilted toward the private side. Therefore, the economic case for full public subsidies is weaker at the college level than it is at the grade-school and high-school levels.

*Skills can also be acquired in other ways: at vocational schools or through on-the-job training, for example. Here we focus on the net addition to skills and income above the gains from those alternatives.

Table 1 Enrollments in public and private schools: High school and college levels, 1930–1985 (in millions of students)

	1950	1960	1965	1970	1975	1980	1985*
High School							
Public	5.7	8.5	11.6	13.3	14.3	13.3	12.1
Nonpublic	7	1.0	1.4	1.3	1.4	1.5	1.4
College							
Public	1.4	1.8	4.0	6.4	8.8	9.1	9.0
Nonpublic	1.3	1.4	2.0	2.2	2.4	2.5	2.4

Source: *U.S. Statistical Abstract*, 1980, p. 143.
* Projected levels.

Actual expenditures on education

These broad conditions are reflected in the actual finances of education. Along with sharp rises in the total spending on education during 1955–1975 has come a major rise in the public share in the educational sector. Table 1 and Figure 1 indicate the scope of this change.

Most education is *public* education, by federal, state, and local sources: Nearly nine tenths of all students go to public grade and high schools. Over two-thirds of college and advanced students go to public institutions.

Public schools for grades 1–12 are financed virtually entirely by tax revenues, as if they were pure social goods. Some schools charge various fees for books and equipment, but these are usually minor. These virtually total subsidies seem to ignore the private benefits that the schooling provides. Other reasons, such as fairness or practicality, apparently explain this complete reliance on public funds.

In higher education, public campuses have grown rapidly since 1960, from 57 to 68 percent of all spending at this level. Enrollment figures in Table 1 show the shift even more sharply. While student ranks in private colleges have only risen by about 500,000 since 1965, they have grown by 5.1 million at public campuses, more than doubling the 1965 level. Two causes are behind this rise. First, there has been a vast boom in capacity, especially at new two-year community colleges. Second, there is the effect of relative prices: Because heavily subsidized public colleges have become much cheaper than less heavily subsidized private colleges, public campuses have absorbed nearly all of the growth in students. In 1980, as Figure 1 shows, while private-college tuition averaged about $3,000 per year, it was only $600 at public colleges and $380 at two-year public campuses. Although the difference is not large compared to the total opportunity cost of college (recall Chapters 2 and 14), it is important enough to decide

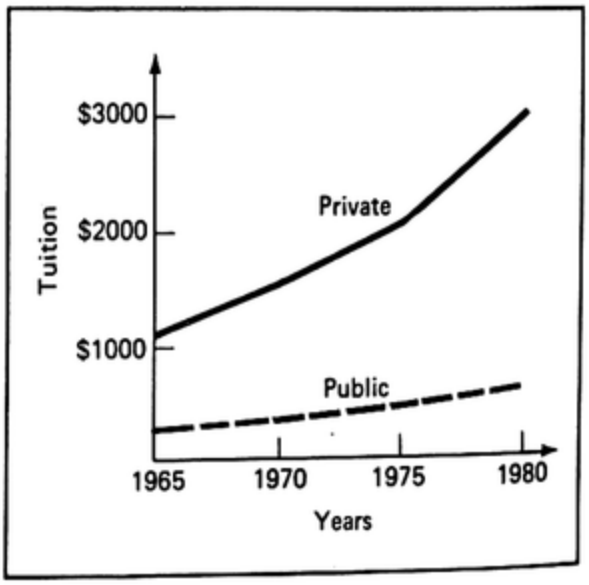


Figure 1 Trends of tuition at private and public colleges and universities

many students' choices. Private colleges have rightly seen subsidized public colleges as a severe form of competition in education markets.

Because the baby boom of 1948–1955 was followed by smaller families during 1960–1968, school enrollments first mounted and then declined. The shrinkage will have major effects on colleges in the 1980s. Enrollments at many campuses had already become smaller during the 1970s, and several private colleges have gone bankrupt. Others may follow as the demand for college education shrinks further. Only as enrollments gradually rise again after about 1990—as is predicted—will education regain its normal condition of moderate growth. In the meantime, you are observing and participating in a system that is in the throes of contraction.

College funds come mainly from government subsidies, as shown in Figure 2. Taking both public and private colleges together, student payments have produced only about 20 percent of all funds. Direct government funds have covered about half of total costs since 1950. Moreover, since private gifts to colleges are tax deductible, the tax revenues of the federal government are reduced by donations to colleges. This provides a "tax expenditure" that benefits college students at the expense of other taxpayers. There is thus extra support for many colleges besides what they obtain from the direct flow of public expenditure.

Moreover, research is important at many larger campuses. On average, it absorbs about 10 percent of college spending. Much of those funds comes as grants from various branches of the federal government. Altogether, over half of higher-education costs are publicly subsidized in these direct and indirect ways.

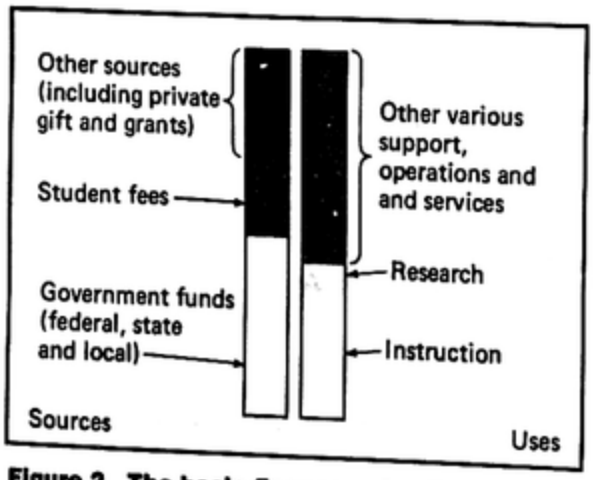


Figure 2 The basic finances of colleges and universities

Among the sources of funds, direct government support (federal, state and local) provides nearly half. Next come student tuition and fees, and then private gifts and grants.

The money goes mainly for instruction. But research also takes a large portion, along with food and buildings. (How does your own campus budget compare with these patterns?)

Public schools and the issue of choice

Because public schooling in grades 1 to 12 is fully subsidized, it has eliminated nearly all competition by private schools. For most students, the local school system is a monopoly, the only available place to go. This gives a mixture of economic effects. There are major ***social benefits*** from having neighborhood-based schools that are available to all youths: cohesion, mutual adjustments among diverse people, and a shared sense of identity and social values. No children are excluded because they lack funds.

But there are also ***social costs.*** Neighborhoods themselves are often ethnic or economic enclaves that inhibit people from mixing with the residents of other such enclaves. The neighborhood school can thus reflect and help perpetuate not diversity but rather the divisions in society.

Moreover, school quality often correlates with neighborhood incomes. Impoverished neighborhoods tend to have poor schools, affluent towns good ones. At the extremes, rich neighborhoods often spend

three or more times per student than poor neighborhoods. Even in the middle ranges, the better schools often have double the resources of the poorer schools. In many poorer school districts, this frequently translates into inferior buildings, inadequate equipment, and a sense of hopelessness. The disparities are caused primarily by the practice of financing schools from local property taxes. The richer neighborhoods naturally can draw on higher tax revenues.

Given those systematic differences in resources, the monopoly nature of the local school systems is important. Though the whole pattern allows for a healthy neighborhood-school cohesiveness, it can also lock children into unequal opportunities. Not only do most students have no choice among schools, but their school usually reflects an underlying economic inequality. Therefore, educational opportunity is often highly unequal, so that poor people have much worse chances for developing their talents. Even if the worst "blackboard jungles" in urban ghettos were remedied, there would remain strong inequalities of opportunity among other neighborhoods.

Two main economic cures have been proposed to provide more equal educational opportunities. *One is to reverse the disparities in resources* by channeling extra resources into schools in poor neighborhoods to offset the inequalities inherent in the neighborhoods themselves. Such specially financed schools could indeed improve students' learning and opportunities, as some practical cases have shown. But the weight of experience is that the political process will not provide the extra funds. Affluent neighborhoods always feel hardpressed just to pay for their own schools; they never think they have a surplus to help pay for other districts. Therefore, the underlying neighborhood inequalities can be expected to prevent a significant shift toward high-quality schools in lower-class areas.

The other proposal is to attack the monopoly aspect of public schools: Let people have free choice among schools, so that consumer preferences can take effect. As a result, poor students could choose good schools in other districts, rather than be confined to their poor local schools. The poor schools would have to improve or else close for lack of students. Each student would be given a "voucher," good for enrolling at any school (public or private). Each voucher would be paid for by the government. The voucher would be no more costly than the present average public expenditure on schooling for children. The best schools would draw excess physical demand—shown by a waiting list of students applying to get in—and would be able to expand, while poor schools would be under financial pressure to improve. New schools could be opened, and, if good enough to draw students, they would offer more choice.

In short, free choice would operate in a setting much like a free competitive market. The public monopoly would be ended, so that families' preferences could take effect. There would be flexibility and variety, as new schools opened where demand was greatest. Fairness would be assured by giving equal-value vouchers to all, so that even the poorest child could try the same schools that are now available to the wealthy.

The approach is based firmly in neoclassical economic concepts of consumer choice, competitive processes, and efficient allocation. Its strongest advocates are classical liberal economists, such as Milton Friedman. With their vigorous endorsement for over 15 years, the method has had some practical testing. Several pilot studies have been done in school districts, with favorable results. Indeed, much of the school-

ing in the Netherlands is already on a similar basis. There new schools can be set up and receive public funds if they can attract enough students. In this setting, Dutch schools actually do offer variety and good quality.

Yet there are limits on the free-choice approach. A voucher is of little use if the only decent school is 20 or more miles away. Even for a three- or four-mile trip, the time and costs of travel could be a large barrier to many poor children, while affluent families could afford to transport their children to the best schools. There would also be severe problems of adjustment in the short run: The best schools would immediately be oversubscribed, so that some method of rationing the excess demand would have to be applied.

Both approaches address a real problem—inquality and monopoly in the public schools—but in opposite ways. The voucher plan had a certain vogue in the 1970s, but that has faded more recently. In general, equal educational opportunity is no longer seen as a major route for social reform in the United States, now that the bright hopes of the 1960s have dimmed. **Economists can agree that the monopoly element of schools aggravates the problems of unfairness and inefficiency in the present system of schooling. Yet, the political process seems more likely to maintain those problems than to solve them.**

Financing public colleges: Efficient? Fair?

Finally, we turn to an important economic feature of college education: whether their financial basis is efficient and fair. Public colleges and universities are highly subsidized: The price (tuition and fees) often covers less than one fourth of total costs. On some public campuses in New York and California, the subsidies are virtually complete. In recent years, efforts to raise the fees to significant levels have caused near-riots on some of those campuses. "Universal education," it was said, would be undermined.

The subsidies pose two main economic questions: Is it efficient to let public college compete on a cut-price basis? Is it fair to channel subsidies from taxpayers to college students' families?

Efficiency If all schools were strictly private, the market outcome would be a variety of schools of differing quality. They would be priced at various levels to cover their costs. For a minimal education, students could try a low-cost local school. For top-quality schooling, they could pay much more at one of the best colleges (if they qualified for admission). On this basis, prices would fit costs at each college, and no campus would be subsidized. The actual array of colleges in the United States shows much of this variety, ranging from small-town junior business colleges to the top public and private universities.

Fairness But there is an important departure from such an efficient market outcome. Public campuses get subsidies, which are often substantial. The economic result is illustrated in Figure 3, for a simplified case with just one standard quality of education. Panel I shows the demand for places at public colleges, given whatever prices are charged by private campuses. The supply of such places, at public colleges, at the prevailing real costs (of teachers' time, buildings, libraries, etc.) is shown by Curve S_1. At a cost-covering price set at P_1, the number of students shown by level Q_1 will choose to attend those public colleges. Meanwhile, in Panel II for private colleges, similar demand and supply curves result in a price of P_A and Q_A number of students.

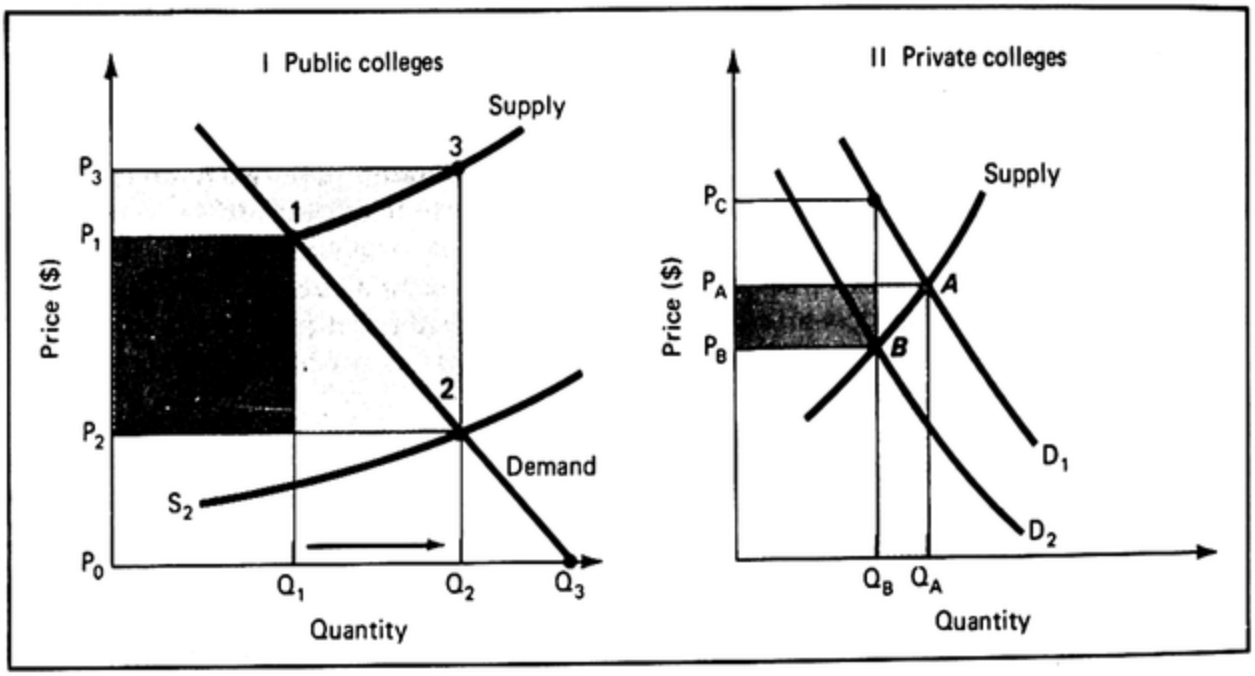


Figure 3 Subsidizing public-college education has mixed benefits

The subsidy shifts down the supply curve in Panel I. Enrollment increases from Q_1 to Q_2. The tuition charge is only P_2, even though the true cost is P_3. (If tuition were cut to zero, with a total subsidy, as many as Q_3 students would enroll.)

The subsidy is the shaded rectangle $P_2P_3$32 (which equals $Q_2 \times P_3 - P_2$). Some of the subsidy goes to make it possible for students from poor families to attend college. But part of the subsidy (shown by the smaller, darker box) goes to students whose families were willing to pay the full original price of P_1. That subsidy is a windfall gain that does nothing to improve educational equality.

Private colleges, meanwhile, face a reduction in demand from D_1 to D_2 in Panel II. At the new equilibrium, fewer students attend even though the price has fallen to P_B.

Now suppose that subsidies are introduced. The supply curve for public colleges shifts down to S_2 in Panel I because the lowered price for public colleges draws students away from private campuses. More students (at Q_2) will choose these bargain-priced campuses because demand has some price elasticity. The students whose families could not afford P_1, but can afford P_2, are those lying between levels Q_2 and Q_1. Those added students are drawn away from many other lines: jobs, vocational courses, and private colleges. The subsidy does make college education more widely available. A total subsidy, at a zero price of P_0, would have brought in even more students, to the level Q_3.

Meanwhile, private colleges now find that their demand curve has shifted inward because the price of a close substitute (namely, public colleges) has been cut, drawing away some of their students. Private-college demand is now D_2 in Panel II, and the new outcome at B gives P_B and Q_B: fewer students and some lowering of price. The cutbacks from A to B reflect the closing of marginal, high-cost colleges. The private colleges maintain that they are faced with unfair competition, for they have much less access to direct public subsidies. They do get some moderate, indirect public support because donations to them are tax deductible. Those "tax expenditures" are not trivial, and they do encourage people and firms to donate more funds to private colleges (and public colleges) than they otherwise would. But that indirect benefit to private colleges falls far short of the direct subsidies provided to public colleges.

Are the subsidies efficient and fair? Economic analysis shows several separate effects, as follows.

CLOSURE OF PRIVATE COLLEGES By forcing some private campuses to close, the subsidies withdraw resources from them whose value was $Q_A - Q_B \times P_A$ (the shaded vertical rectangle in Panel II of Figure 3) *plus* $P_A - P_B \times Q_B$ (the horizontal shaded rectangle). At Q_B, the original customers of private colleges thought that the value of the colleges' services was as high as P_C, as shown. Some of those students have now migrated to public campuses. But the disparity between P_B and P_A indicates that cost is now out of line with value at the margin (the margin being level Q_B).

FINANCIAL PRESSURE ON PRIVATE COLLEGES Since the remaining private colleges now take in only a revenue (or price) of P_B per student, their funding is reduced. Their total revenues are $P_B \times Q_B$ which is much smaller than their former revenues of $Q_A \times P_A$. They will have to reduce costs, and possibly services and quality, too.

AT PUBLIC COLLEGES, VALUE AND TRUE COST ARE SEPARATED The true cost at public campuses is now P_3 on Supply Curve S_1 (the supply curve still reflects opportunity costs, even though subsidies have altered the price paid by students). P_3 is the private payment per student (P_2) plus the subsidy ($P_3 - P_2$): Together they just cover the average total cost for the education of Q_2 students.

Note that the marginal value of public-campus education is now down to P_2, reflecting people's preferences and their ability to pay. Meanwhile, the marginal cost of education is much higher (at P_3) than the marginal value. This large deviation between price and marginal cost—between value and sacrifice at the margin—reflects an inefficient allocation. Students are receiving services that cost three times as much as they are valued at the margin.

AN UNNECESSARY SUBSIDY But the cost of this result is a large subsidy, whose total amount is shown by the shaded rectangle $P_2P_3$32. That amount reaches into many billions of dollars. Much of that subsidy goes to unneedy students' families, who were already able and willing to pay the original price of P_1 to the public college. Since those Q_1 students now only have to pay the bargain price of P_2, they get a subsidy shown by the shaded rectangle $P_2 - P_1 \times Q_1$. That functionless subsidy simply goes into their pockets without changing any of the Q_1 students' choices. In Figure 3, the wasted subsidy is nearly half of the total subsidy paid to the public colleges, as you can see by comparing the two shaded areas.

To reach a judgment on the whole matter, economists would consider efficiency first and then equity. *Efficient allocation is disturbed by driving the wedge between price and marginal cost.* The deviation is not small: "Too many" students are drawn to public campuses, compared to the costs of serving them and to the best alternative uses of their time. And private campuses are cut back from their efficient levels.

Some amount of "dynamic" efficiency may be recouped, however, if the added public-college students (between Q_1 and Q_2) become much more productive than they would otherwise have been. Their later productivity could offset some of the distortion that is caused in the current allocation. How large these relative benefits might be is not known.

The equity effects are much clearer. The subsidy is large and indiscriminate, going to all public-college students whether needy or not.

Economists point out that scholarship programs based on need would be more ef-

ficient and fair than the present broad subsidies given to public colleges. Tuitions would be set to cover the full cost of education, but scholarships would be provided fully—up to the direct costs of college *and* forgone income—for poor students who are qualified and can show definite need. The scholarship aid would flow only where it would provide equal educational opportunity. It would avoid the current inefficiency and functionless subsidies, while getting all of the possible gains from giving access to college for poor but talented students. And the cost of scholarships would be a small fraction of the present subsidies.

Despite this clear analysis, the traditional subsidies continue. Yet, there has recently been some revision toward better patterns, because the taxpayers' willingness to pay for education receded in the 1970s. The result has been a moderate shift toward higher tuitions and larger scholarship programs in some states. But this shift is short of what microeconomic analysis calls for.

Social regulation: Protecting the environment, workers, and consumers

Since 1960, there has been a rapid growth in two kinds of social regulation: abating pollution and protecting the safety of workers and consumers.* Because these policies have costs and effects, they raise economic issues. Amid much criticism, the policies have made some progress. We present them together here, for they involve the same basic kinds of ***cost-benefit analysis.*** First, we take up environmental issues. Then we discuss worker and consumer protection.

*They differ from the "economic regulation" that we presented in Chapter 13. There, the regulation controls prices, profits, and the ability of firms to enter and compete in markets. The aims are to prevent monopoly's bad effects while obtaining economies of scale.

Social regulation, by contrast, deals only with *social* aspects of production: pollution and safety, at work and in the design and use of consumer products.

Environmental issues and programs

The environment can be harmed in many ways: the pollution of water and air, the destruction of wilderness and underground water sources, the infliction of loud noises in cities. Industrial factories often cause the damage, but normal life in crowded cities also creates numberless external effects. Correcting those problems is a leading area for political action, as Chapter 18 has noted.

Until the 1960s, most economists regarded such environmental problems as minor and of no lasting impact. The few serious ones—such as polluted rivers, and smog in cities such as Pittsburgh (where street lights were often kept lit all day through the 1940s) and Los Angeles—were regarded as a regrettable but small price to pay for economic progress. Efforts to prevent or cure the pollution were said to be too costly and likely to cause barriers to industrial growth. Business interests held more sway than in the 1970s and 1980s, and had little interest in limiting themselves to reduce pollution.

When the environmental movement took hold in the later 1960s, economists developed the analysis of the economic causes of pollution and showed how to design efficient cures. The economic remedies are reasonable, simple, and clear for many pollution problems, but the political problems are often quite complex. Accordingly, many policies have been inadequate or unnecessarily costly.

Over $500 billion has been spent since 1968 in cleaning up pollution in the country. The yearly totals of cleanup costs have

risen from about $25 billion in 1972 to $61 billion in 1981. A further $619 billion of costs are projected for between 1982 and 1988 (in 1981 dollars), even with no new laws, according to estimates by the Council on Environmental Quality. These cleanup costs have been estimated to raise the inflation rate by 0.2 percent, to raise unemployment rates by 0.2–0.4 percent, and to slow national productivity growth by about 10 percent. Because large costs are involved, the United States may face difficult choices between the quality of life and economic growth.

Policies have been set mainly by the Environmental Protection Agency (EPA), which was created in 1969. The EPA has had only modest powers and funds and has been slow to develop effective tools. The main methods available to it are summarized in Table 2, with examples. Despite much criticism, both fair and unfair, the EPA has reduced some kinds of pollution. Its techniques have also improved, moving gradually from just issuing rules toward using economic incentives. Certain states have also taken specific cleanup actions. Congress has provided various financial subsidies, trying to encourage pollution abatement. These actions have brought improvements in some types of pollution and in some areas.

Yet, much pollution remains, as Figure 4 indicates. Large areas of the country still have severe air and water pollution. Acid rain—caused by sulfur dioxide and nitrogen dioxide—has increasingly sterilized lakes, harmed forests, and contaminated groundwater. Over 30 cancer-causing air pollutants have not been regulated at all.

In short, the cleanup process has been expensive, of doubtful effect, and incomplete. It is important to make further efforts efficient, lest their cost become astronomical. But many pollution-control programs have been criticized for being wasteful. The policies have relied mainly on rules rather than cost-benefit analysis, and microeconomists have been among their sharpest critics. During 1977–1982, some EPA programs began to incorporate economic incentives more effectively.

Pollution control, therefore, poses sophisticated economic issues, which are evolving under debate in hundreds of in-

Table 2 ***Alternative solutions to pollution***

Types	Examples
Persuasion	
Public appeals	Requests for reducing pollution. Exposés of actual pollution. Threats to enact laws.
Rules and Fines	
Sets permissible levels; imposes fines for violation	Equipment standards limiting factory effluents. EPA standards for automobile emissions. Rules requiring "scrubbers" on electric-utility smokestacks.
Direct Incentives: Taxes and Subsidies	U.S. subsidies to cities for better sewage-treatment capacity.
Market-type Solutions	"Bubble" treatments. Marketable pollution permits.

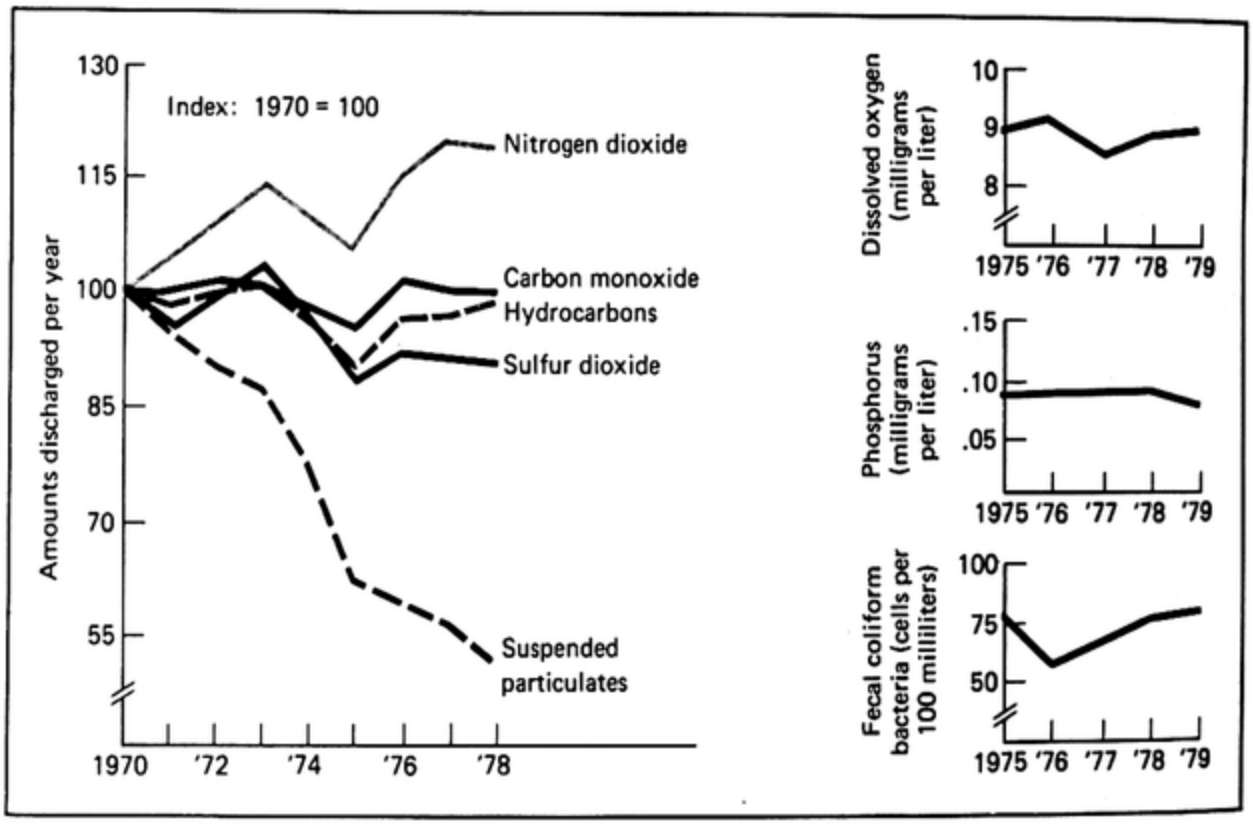


Figure 4 Changes in and persistence of pollution
Source: *U.S. Statistical Abstract.*

dustries and thousands of locations. Amid the variety of pollutions and remedies, several main points of economic analysis are now well agreed upon.

The main problems lie in (1) defining the criteria that should guide the policies; and (2) designing actual policies that achieve these goals efficiently, with a maximum benefit from given costs.

Cost-benefit issues

Economists insist that cost-benefit analysis is the best basic framework for analyzing pollution and its cures. The main economic elements are usually of the cost-benefit kind that is illustrated in Figure 5. We now present those issues in more depth.

The level of cleanliness (of avoiding pollution) is shown on the horizontal axis, with rising levels toward the *right*. To stop pollution means to move to the right, toward a zero level of pollution. The benefits of doing so include the lives that are saved and the health damage that is avoided (recall Chapter 18). Businesses also often benefit from having better water and air resources available to use in production. Thus, a wood-pulp mill might require large amounts of clean water. If only polluted water is available, the mill will have to build a filter plant, at some cost. The extra cost would raise its total cost of production. Therefore, having clean water is a direct economic benefit.

In Figure 5, we assume that these benefit values can be measured reliably, to compare with the costs of reducing pollution. We also assume that those costs are minimized for each level of cleanup achieved, by applying the best pollution-

abatement methods. The efficient control of pollution occurs at Point *A*, where the marginal benefits just equal the marginal costs. The logic is clear because departing from Point *A* is demonstrably inefficient.

A serious problem arises if the costs are paid by one group of people but the benefits go to another group. Then the issue of efficiency is mixed with the question of fairness. This problem is avoided if costs are assigned to the beneficiaries, either precisely or approximately. Where this is not done, cost-benefit analysis has less validity as a method.

Some observers also criticize cost-benefit analysis for allegedly understating the health benefits of clean air and water. The benefits usually involve values of people's lives based on their productivity; that productivity, in turn, is measured by their incomes. Thus, a 50-year-old person now earning $60,000 per year would be projected to earn at that rate for 15 more years, for a lifetime value of 15 × $60,000 = $900,000. The present value of this income would, of course, be lower after discounting future incomes.

This method for valuing lives has limitations, however. *First*, it assigns a zero value to a person who works at home rather than at a paid outside job; the true value of his or her work is ignored. *Second*, it assigns a low or zero value to retired people and young children because they do not hold paying jobs. And *third*, its income criterion assigns low values to janitors and high values to executives: Though possibly efficient, it conflicts with the ethical concept that all people's lives are inherently of equal value.

Past government decisions have frequently set low values on human lives, often in the range of just $20,000 to $40,000. Yet the problem is a *matter of degree*, not

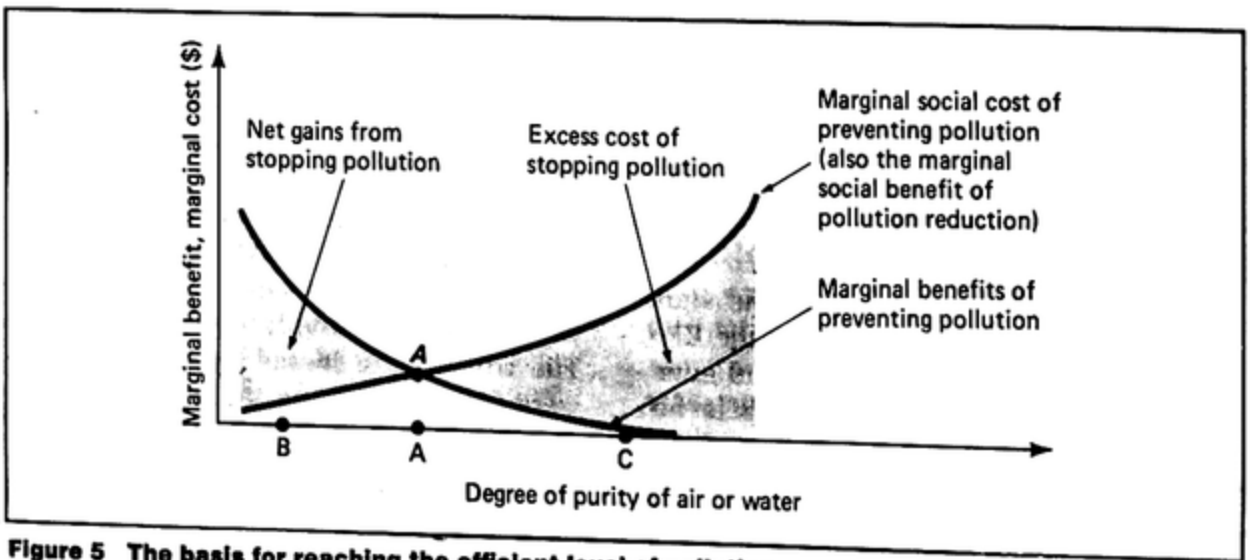


Figure 5 The basis for reaching the efficient level of pollution

Small amounts of pollution cause little social harm because they are assimilated by nature and do not build up in people's bodies. But rising pollution (to the left) causes a rising marginal degree of harm, as annoyance turns to sickness and eventually to death from high pollution.

Meanwhile, the marginal costs of reducing pollution become extremely high to the right, in trying to eliminate the last traces of pollution. But at the left-hand end, the marginal costs are low because the worst degrees of pollution can be reduced relatively easily.

At point *A*, the marginal costs of stopping pollution are just equal to the marginal social gains. Pollution levels above *A* are too expensive to correct efficiently. Pollution levels below *A* cause more harm than it would cost to prevent them.

of logic. The need is for better estimates, rather than for rejecting the cost-benefit approach altogether.

Indeed, human lives are frequently valued in private insurance cases, where settlements have commonly been in the range of $250,000 to $800,000, depending on age and occupation. Such values indicate the order of magnitude that can usually be assigned to saving lives by regulating pollution and safety.

In sum, any rational policy will have to compare costs and benefits. It will (1) include all costs and benefits fully, rather than omitting or understating some of them; (2) weigh all risks properly (risks of harm to people, firms, etc.); and (3) make sure that future interests are properly weighted. Doing that can satisfy the main criticisms of cost-benefit analysis. For many purposes, the debate is primarily about the categories and the amounts, not the basic logic.

Now we turn to specific policies to reduce pollution. They have mainly involved physical rules and limits, which we will discuss first. Only recently have economic incentives been explicitly introduced into the policies.

The use of rules to limit pollution

Congress has largely controlled the standards of clean air and water that the EPA has tried to enforce for factories and automobiles. Timetables have set targets for successive years. Costs are given little standing as guidelines for EPA decisions. Instead, the EPA has relied extensively on the following approach: The "best available control technology" is required in new equipment, as long as it is "feasible"—that is, its costs will not put a large share of the industry out of business.

These two methods—permitted levels of pollution and "best (feasible) technology"—can fit cost-benefit outcomes but only by chance. Instead, they may go to either extreme, Even where they happen to hit the right degree of purity, these methods lack proper incentives for firms to reduce pollution efficiently.

Moreover, the controls have mainly applied to *new* equipment, such as machinery and automobiles. That has (1) too sharply raised the cost of new equipment, discouraging innovation and reducing productivity; and (2) ignored possible major gains from reconditioning old equipment and automobiles.

The EPA has relied on setting physical standards for air and water quality in geographic areas. Factories are then required to reduce their emissions by certain percentages. They must install low-emission equipment in new factories. In areas already meeting the standard, complex rules permit only certain "increments" of new pollution. The rules and procedures are often time-consuming. Two cases—automobile emissions and the steel industry—illustrate the main economic issues.

Automobiles In 1970, Congress set a schedule of automobile emission standards to be met in a detailed timetable of successive years. The main harmful emissions were to be reduced by 90 percent by 1975. The benefits were not measured precisely; 90 percent was simply set as the target. The true harms caused by the emissions were not known, even approximately, and they were still largely unresearched.

In practice, the targets have frequently been moved back, though by 1981, a large share of harmful auto emissions had been eliminated. Two features of this program were especially debatable on economic grounds. One was the rush to require costly catalytic converters on U.S.-made cars after 1974. Cheaper methods for reducing emissions were developed by Japanese automobile companies at much lower

cost, using many small changes in engine design. Moreover, a large proportion of the catalytic converters on U.S. cars have been disconnected or ruined by misuse, while the Japanese cars could not be so easily changed to permit emissions.

The costs of emission controls reached $40–$110 per year for 1981 model cars. Did the benefits justify the costs?

In appraising the benefits and costs of emission controls, one crucial point is that automobile pollution is not evenly spread. Its harmful concentrations occur only in certain large cities, while virtually all auto emissions elsewhere—more than half of total emissions that occur—are dispersed harmlessly in small towns and open spaces. Therefore, at least half of emission-control costs are unnecessary.

Conversely, the existing controls still leave pollution levels relatively high in the main large cities. If pollution-dense zones could be defined, then the controls could be confined to cars operating in them. The rest could be free of the extra costs. Or possibly there could be pollution taxes on cars operating in high-density urban areas.

Yet, actual policy has instead forced *all* new cars—and *only new cars*—to meet engineering emission-control standards. The alternate, more precise approach was regarded as impractical, since cars move freely among areas. Moreover, inspection programs have not been implemented, even though they could cheaply identify the worst polluters and enforce repairs.

The United States has stopped a large share of automobile emissions, but in inefficient ways. If pollution is to be reduced further, it is important to apply more efficient methods.

Steel In recent years, the steel industry has made good progress in complying with federal pollution standards. As of 1981, some 87 percent of steel plants complied with pollution standards (in most other industries, compliance is between 10 and 25 percent).

The EPA has relied mainly on rules and fines in a series of negotiations with this industry to press the companies to invest in cleaner technology. Industry officials say that the new equipment adds 25 percent to the volume of its investment needs, even though the average cost of steel is raised by less than 3 percent. Since imported Japanese steel is now available and competing strongly at lower prices than many U.S. firms can match, the U.S. companies find it natural to blame pollution control for the loss of steel production and jobs. "Cleaner air costs jobs" is the apparent dilemma.

But the real issues go deeper. For decades, the industry has been so sluggish in adopting new technology, that U.S. firms are now paying Japanese firms to show them how to improve their efficiency. If the U.S. industry had been more innovative and efficient during 1930–1970, the current cleanup could easily be funded while still letting the industry meet Japanese import competition. (The Japanese firms already meet high standards of pollution control in Japan.) Because of the U.S. industry's failures, the EPA rules now do seem to impose difficult choices on steel factories in some parts of the country. One solution might be to provide public funds to help finance the pollution-control equipment, along with other actions to rehabilitate the industry.

In short, this case involves more than simple rules or subsidies. Pollution control needs to be included in a larger approach to industrial innovation in the steel industry.

The use of incentives

Though many economists have urged that more ***economic incentives*** be applied, little of that has yet been done. There is increas-

ing experimentation with three specific methods: the "bubble" concept, "marketable permits," and taxes on emissions.

The bubble concept The EPA has begun applying a "bubble" concept in some cases since 1977. This approach applies to firms that have several or more pollution-emitting plants in an area which is treated as a single unit or "bubble." They are allowed to adjust among those sources of pollution, as long as their total emissions are within permitted levels. The method lets firms decide how best to reduce their emissions. Otherwise, the agency would have to reduce emissions from each factory, stack by stack, which would make it harder for each firm to design efficient ways of reducing its total pollution. By 1981, over 80 "bubbles" had been established, and the practice was spreading rapidly.

Marketable permits Some economists regard marketable permits as even more promising. They are specific emissions rights, which can be bought and sold. They work as follows: Suppose a 33 percent reduction is deemed appropriate, based on a full cost-benefit analysis, in an area with 150 factories that pour out 15,000 tons of pollution daily. The agency first establishes severe fines for firms that pollute at levels higher than their permits. Next it prints permits for 10,000 tons daily. It then either (1) gives or sells these to the existing polluters in proportion to their emissions; or (2) gives them to citizen groups or cities in the area that suffer from the pollution; or (3) sells them off to all comers at auction.

If enforcement is complete, pollution should drop to the 10,000-tons-a-day rate, but the 150 firms could adjust by the least-cost methods. Complex, burdensome rules would be avoided. If method 2 were chosen, the factory owners would have to pay to the sufferers of pollution a dollar amount that approximates the social costs that the pollution inflicts. Therefore, this method would come closest to a complete economic treatment of pollution by making the *external* costs of pollution an *internal* cost to the polluters. Moreover, the victims of pollution would be compensated for their losses.

Under methods 2 and 3, citizen groups or government units could buy and reserve some of the rights, thereby forcing even lower levels of pollution. This would allow people to cut the pollution they dislike directly. It would also force the foes of pollution to consider the real costs of the cleanup, by testing what they are willing to pay. (But note that poor people who suffer from pollution could not afford to purchase pollution rights.)

If marketable permits are accurately priced and backed by strict enforcement, the results are close to an economic ideal because they let decisions be controlled by marginal costs and benefits. Those who know best (managers) and care most (citizens groups) could act without relying on a bureaucracy.

Taxes on emissions are a more direct method and one also favored by many economists. Such "effluent fees" are simple to use and apply economic incentives to reduce pollution. The agency's main task is to set the tax at a level that will reach the right outcome. In Figure 6, the tax would be set at T_A per unit of pollution, which equals both the marginal cost and the marginal benefits of pollution control in equilibrium. That would make the external cost of pollution an internal cost to the polluting firm. The firm would then be willing to spend up to that amount to reduce the pollution. Therefore, firms would respond by reducing pollution to level P_A, making their own best choices about how to do so most efficiently.

The logic of these several methods is clear, but putting them into practice could

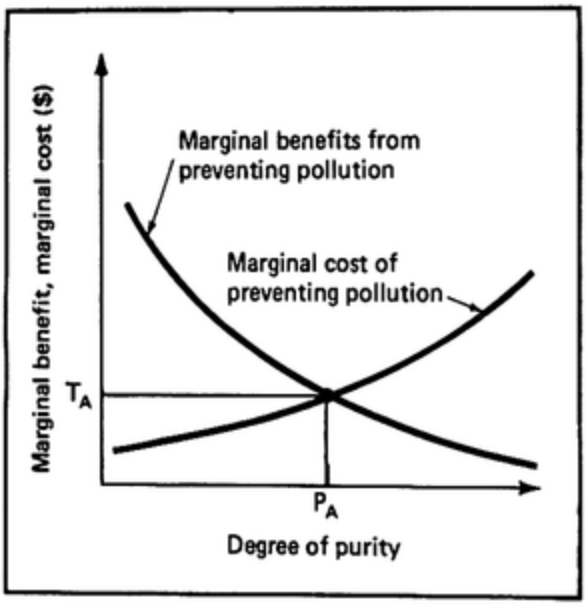


Figure 6 Setting an efficient pollution tax

The efficient level of pollution is at P_A in this illustration, because that is where marginal costs and benefits are equal. A tax set at the value T_A would then confront the polluters with the social cost of their actions: The external cost would become an internal cost to the firms. They would then have the incentive to spend up to that amount so as to avoid the tax. They could do so in the most efficient way.

be complex. Since the marginal values are rarely precisely measurable, some guesswork is necessary. Because the numbers are debatable, companies resist, and the taxes, fines, and enforcement could become entangled in debates and lawsuits.

Nonetheless, each of these methods can improve on the rigid rules of the 1970s. Pollution taxes are especially appropriate for emissions that cause widespread, long-distance problems like acid rain. "Bubbles" and marketable permits are better suited to localized pollution problems. Though precise optimal outcomes are difficult to predict, these methods do give strong incentives, with little bureaucratic involvement. That fits the economists' usual preference for price signals rather than regulations.

In any event, these three methods require monitoring the companies' pollution levels. When companies are few, these costs will be low. But when hundreds or thousands of companies are involved, the monitoring costs may overwhelm the economic advantages of these incentive schemes. Therefore, the best practical control of pollution will vary from industry to industry. These choices, too, can be guided by comparing the costs and benefits of alternatives.

Programs protecting worker and consumer safety

Since 1970, three federal agencies have been created to promote safety in the workplace and in consumer products (joining the Food and Drug Administration (FDA), which was created in 1906). As summarized in Table 3, the new agencies are a response to real problems. Over 100,000

Table 3 New federal agencies for worker and consumer protection

Agency (year created)	Budget 1981–1982 ($ million)	Staff Members	Formal Purpose
Occupational Safety and Health Administration (OSHA) (1970)	$193	3,260	To protect safety and health on the job
National Highway Traffic Safety Administration (NHSA) (1970)	$141	825	To reduce traffic accidents, by requiring better design of motor vehicles
Consumer Product Safety Commission (CPSC) (1972)	$58	899	To reduce product-related injuries to consumers

people are killed each year by industrial, consumer, and other accidents; over 5 million people a year undergo injuries that disable them for more than one day. The costs in treatment, lost work, and suffering are large. Millions of workers are still exposed to significant risks of accidents and such health hazards as lead, vinyl chloride, cotton dust, coal dust, and arsenic. Many consumer products also are hazardous. From automobiles, motorcycles, and bicycles to foods, medicines, poisons, ladders, cribs, and children's toys, there is a variety of products that can harm people. The question is how much public protection is needed.

Like the FDA earlier, OSHA, CPSC, and NHSA have had the usual history: Created to solve large problems, the agencies have grown rapidly, made some mistakes, stirred opposition, and, since 1980, been trimmed back. Often criticized as harmful bureaucracies, they have experimented with a variety of methods, just as the EPA has. Economists have urged them to use cost-benefit criteria. Instead, they have often sought to provide extreme, absolute degrees of protection that deliberately ignore costs.

Costs and benefits The basic economic issues are identical to those involved in protecting the environment, as illustrated in Figure 5. But it is the degree of safety from bodily harm that is measured on the horizontal axis in Figure 7. Providing safety incurs economic costs, which rise at the margin. Reducing the worst hazards gives high benefits, but attaining near-absolute protection often gives only small marginal benefits. The valuation of the benefits from safety is done as before, by estimating the benefits of avoiding the suffering, lost production, and other impacts. For example, a worker who loses his or her legs not only suffers personally, but also must pay for artificial legs, wheelchairs, and other assistance.

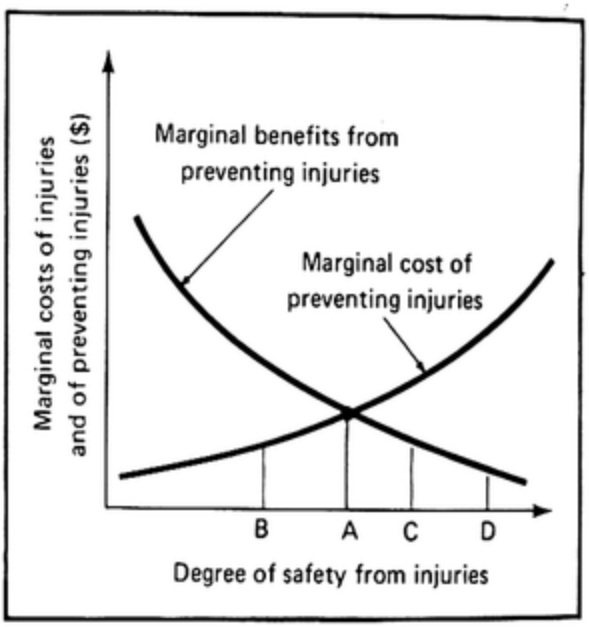


Figure 7 Marginal costs of injuries and of preventing injuries

The conditions correspond to those in Figure 5. *A* is the efficient level of safety protection. *B* is too little protection, while *D* and *C* are too much, by cost-benefit criteria.

The marginal costs of protection rise, as shown in Figure 7. Cheap protection such as guardrails, hard hats, goggles, and the like may sharply reduce risks, as at Point *B*. But reducing risk down to very low levels, as at Point *C*, may require that machines and factory layout be redesigned. Therefore, an effort to provide, absolute safety, like absolutely clean air or water, can be exceedingly costly.

The efficient degree of-risk is at Point *A* in Figure 7. Ideally, OSHA, CPSC, and NHSA would design policies to reach that point in each case they handle. For example, NHSA would require only this amount of crash protection (seat belts, bumpers, reinforcements, etc.) in cars. Similarly, the FDA will require foods to meet only these marginal conditions of safety, and if a guardrail near steel furnaces fits Point *A*, OSHA will not require more protection than that.

Yet the laws creating FDA, OSHA, and CPSC stressed the *benefits* of worker and consumer safety but gave little attention to *costs*. For foods, *any* risk of causing cancer would require a ban. For OSHA and CPSC, the greatest "feasible" protection was mandated; cost was a limit only if it would be large enough to force many or most firms out of business. Later Supreme Court rulings have upheld this willingness to enforce safety even if it imposes costs that exceed benefits. Thus, the laws appear to call for Point *C* or *D*, if that is necessary.

In practice, the agencies have usually followed that priority. They often make no explicit effort to locate the efficient degree of protection corresponding to Point *A*. Instead, OSHA, for example, requires that all producers adopt the safest technology now available. It has also set levels for toxic fumes and dusts—such as benzene, asbestos, arsenic, lead, and cotton dust—that require changes in machinery. These standards have been accused of being too tight by cost-benefit standards.

Cotton dust is a good example. It causes byssinosis (brown lung disease) in cotton-mill workers. When it is present at levels of about 200 micrograms per cubic meter, about 13 percent of workers will contract the disease. In 1978, OSHA set limits on cotton dust, requiring firms to reduce the levels substantially. These standards mean that it will cost about $135,000 to avoid each case of brown lung disease. The benefits have been calculated at only about $20,000 per case, suggesting that the rule is too strict. Alternatively, the same degree of protection could have been reached much more cheaply by requiring employers to pay workers at least $20,000 (or up to $135,000) if they contract the disease. The Reagan administration sought in 1981 to relax standards such as these, in line with their new cost-benefit emphasis. But through 1982, the Supreme Court has upheld the stricter approach. Consumer safety protection by CPSC has followed similar lines, with no close reliance on cost-benefit comparisons.

Military spending

The U.S. military system is the biggest single spending program in the country. Figure 8 shows its size and main trends. Much military spending goes to pay for personnel: officers, soldiers of all ranks and skills, and people in many other job categories. The rest goes to buy weapons and materials, ranging from pencils to complete missile systems. These items now make up over 70 percent of all purchases by the government.

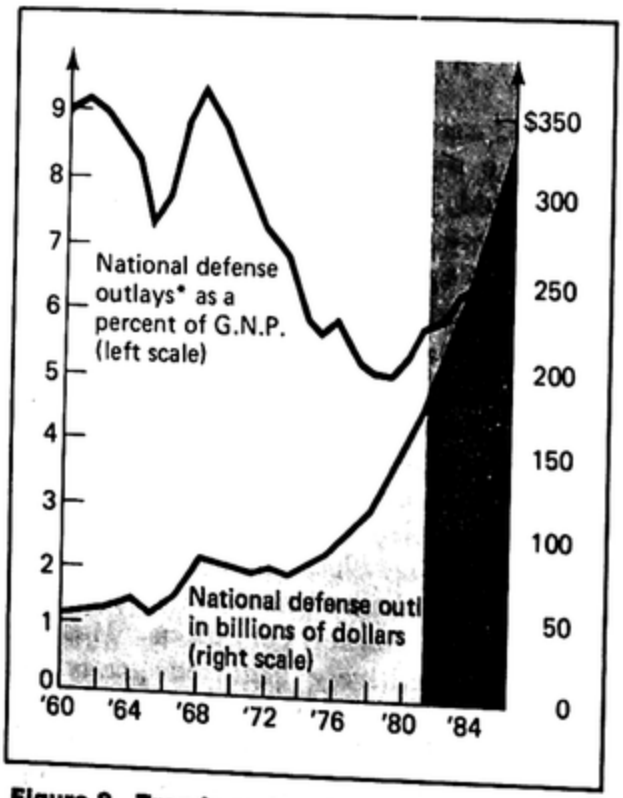


Figure 8 Trends and projections of U.S. military spending

*National Defense includes outlays by the Defense Department, the Energy Department, the General Services Administration, and the Selective Service System.

Source: Office of Management and Budget.

The Pentagon is also a large owner of capital. Besides its own bases, inventories of armaments, and transportation networks, it owns 149 manufacturing plants and over 100,000 items of industrial equipment. Military spending also plays a large role in total U.S. research and development (R&D). It pays for over half of all R&D carried out in the United States, and in certain key industries—aerospace, aircraft, and related technology—it buys nearly all of the R&D that is done.

This huge flow of money for arms, people, and technology poses three main economic issues: (1) How can waste be avoided in producing the supplies? (2) How much military spending is appropriate, neither too little nor too much? (3) How can a volunteer military force be arranged, so as to avoid the costs of a military draft? We will treat each issue in turn.

Avoiding waste in producing military goods

Wartime procurement always entails urgency and, therefore, an unavoidable element of waste. In peacetime, however, the Pentagon could ensure that its materials are supplied efficiently, at minimum cost. Like conventional buyers in private markets, it would seek the lowest-price sellers for each item. The goods would be made at minimum cost, and the price for each would cover only that cost, including the cost of capital.

In practice, less than 10 percent of military purchases are made under such competitive market conditions. Most of those competitively acquired goods are simple ones like foods, clothing, and furniture. The other 90+ percent are bought after considering only a few suppliers and picking one. The terms are negotiated rather than set by impersonal markets, and often the military services end up paying much more than they originally contract for.

In sum, most weapons purchases are made outside the standard conditions of efficient competitive markets. The main deviations are shown in Table 4. They reflect an unusual set of conditions found in military weaponry. Each new weapon (e.g., an aircraft, tank, ship, warning system) requires new designs and testing. In developing the plans, the Pentagon officers usually deal with only a few firms. Those firms may compete strenuously to win the crucial first contract. But the winner anticipates having the whole production run—of M1 tanks, B1 bombers, missiles, Trident submarines, or whatever. The supplier-buyer relation tends to become fixed and intimate, as in any bilateral monopoly or oligopoly.

At any time, the whole flow of complex armaments involves scores of such situations. They evolve and change, but the Pentagon and the main suppliers continue in a pattern of working relationships. The leading 20 suppliers in 1967 through 1977 remained mostly the same. President Eisenhower called this pattern of interests the "military-industrial complex." Retiring military officers frequently take jobs with the companies, and company leaders are often appointed to high policy positions in the Pentagon.

This complicated set of relationships between firms and military officials gives some genuine social advantages. Officers know the companies well. Plans are made under conditions of close cooperation. The companies are supported in business, so that they are ready to meet wartime emergencies. There is a high degree of continuity and mutual understanding.

But there are also economic and social costs. The planning of new weapons is often imbued with optimism, as company officials anticipate and offer to the military officers the kinds of weapons that fit their preferences. The focus is often on getting the highest technical quality of weapons,

Table 4 ***Contrasts between conditions in competitive markets and many weapons markets***

Competitive Markets	Many Weapons Markets
Many small buyers	One buyer (DOD)
Many small suppliers	Very few large suppliers of a given item
All items small, perfectly divisible, and in large quantities	One ship built every few years, for hundreds of millions of dollars
Free movement in and out of the market	Extensive barriers to entry and exit
Prices set in line with marginal costs	Prices proportional to total costs
Prices set in line with marginal utility	Any price paid for the desired military performance
Market shifts rapidly to changes in supply and demand	7–10 years to develop a new system, then 3–5 years to produce it
No government involvement	Government is regulator, specifier, banker, judge of claims, etc.
Selection based on price	Selection often based on politics, or sole source, or "negotiation"; only 8 percent of dollars awarded on price competition

Adapted from J. S. Gansler, *The Defense Industry*. Copyright © 1980 by Jacques S. Gansler. Used by permission of the MIT Press, Cambridge, MA.

with little regard for cost. As new weapons develop from mere ideas to early designs, both the firms and the military become committed to the need for producing the new weapon.

Meanwhile, the hopeful companies make ambitious proposals. To "buy in" and get the first contract, they often make deliberately low bids, even though they expect to have to ask for higher payments later as production occurs at higher costs. Their *marginal* cost for new orders is low when they have a small backlog of orders, and they need only cover *variable* costs in the short run. Yet, the contract payments must eventually cover the companies' average *total* costs.

So each supplier has an incentive to set low prices in its bids when it is short of orders. Later, when production is under way and costs turn out to be higher than predicted, the military has only the one supplier. It then often must acquiesce in the cost overrun and cover the extra costs. But that overrun will frequently be ascribed to changes in the weapon's design and to the weapon's high technical quality.

Most of the buying decisions are made by middle-level officers trained as engineers, who know little about economic efficiency. The contracts are usually on a "cost-plus" basis: The supplier is paid for its production costs plus a fee for its efforts. Thus, for a 300 fighter-plane order, the military may pay $2.4 billion ($8 million apiece) for the $2.1 billion of costs plus a $300 million contractor's fee for carrying out the production. If costs had been $1.5 billion or $3.5 billion, the $300 million fee would still have been the same.

In such "cost-plus" contracts, the supplier has no incentive to minimize costs once the contract is signed. Some contracts even tie the size of the fee to the level of the costs themselves, so that larger costs yield a larger fee. In these cases, the contractor is directly induced to increase the use of resources by raising the level of the costs. But even a neutral "cost-plus, fixed-fee" contract departs from sound private-market standards.

Such cost-plus contracting has long been recognized as inefficient, but it remains standard for the Pentagon. The weapons are said to be too new, and complex for a price to be specified in advance. And the initial competition for the contract, which forces the winner to offer a low bid, is itself a second main cause of inefficient allocation, for it makes it seem that the weapon will be much cheaper than it inevitably turns out to be. Many current weapons systems would not have been chosen if their true costs had been known in advance. When a firm offers to build for $1.9 billion a nuclear warship that will eventually cost $4.5 billion, it encourages the Navy to purchase the ship. The military can refuse to pay the higher costs when the production is completed, but often the supplier can argue that important new features were added to the weapon's design by the Pentagon, so that the firm should not be responsible for the excess costs. If the military officials do force the supplier to undergo large financial losses on such contracts, they may bankrupt the company and lose its capacity to produce weapons in the future.

This economic process continues, generating new families of weapons systems, which frequently cost over 50 percent more—and occasionally over 200 percent more—than was planned. In each case, there is argument about the true cause of the cost increase; each case has its unique features. But there is no doubt that an element of cost inflation exists.

Pentagon officials concede that much inefficiency occurs. The best economic estimates are that the waste is in the range of 10–20 percent of total military purchasing. That would be about $7–$15 billion yearly in recent years. As a contributory factor, the military leases over $200 billion of buildings and equipment to suppliers at virtually no cost. In this setting, the military preference for high technical reliability rather than minimum cost inevitably leads to significant excess costs.

The main economic causes of the problem have been well known and studied for years, but a cure would require reorienting the basic thinking and techniques of the thousands of officers responsible for the military's purchases. Though economists may continue to argue for more efficiency, there is little probability of change.

Efficient military levels and the arms race

Military expenditures take a significant share of total world production. At over $650 billion, the total spending on armed forces in 1981 was 6 percent of worldwide economic activity. In the same year, the international trade in weapons was $120 billion. This accelerating global arms race absorbs large flows of scarce resources and raises the risks and impacts of warfare, with its severe economic disruptions.

The economics of the arms race derives from both budgetary overstatement and oligopolistic competition. We discuss them in turn.

Military estimates Economists and others have long discussed the tendency for the military to overstate its economic needs. To provide military security, the Joint Chiefs of Staff naturally prefer a big army and navy and a large stock of weapons, ca-

pable of meeting all likely threats. Each report of the Soviets making better planes, tanks, submarines, and missiles raises pressure to provide the same to our military services. To attain absolute security, we would have to devote all national production to military ends, and that still might not be enough.

For economists, military needs are a matter of degree. An efficient policy would provide the total of weapons at which the marginal benefits (in enhanced national security) equal the marginal costs. Within that total, each weapons system and type would also be provided just up to the efficient margin.

Therefore, trade-offs have to be made at two levels: between military and civilian needs, and among military choices (such as submarines, B1 bombers, M16 rifles, and footsoldiers). The natural role of military officials is to request everything they might need, and that of the government to reduce the levels to those consistent with efficient total allocation.

Ideally, those optimum marginal conditions could be identified and the military budget nicely calculated in total and in all its parts. In practice, the benefits of weapons and personnel are mostly matters of conjecture. Admittedly, some direct calculations can be made. For example, given the known Soviet armored capacities, how many tanks, artillery, and other forces would be needed to "win" a certain type of war in Europe?

These and other practical "scenarios" for prospective military actions are modeled repeatedly by Pentagon planners, using many assumptions. The analyses give some practical estimates about the weaponry needed to achieve specific objectives. Moreover, data about Russian weapons budgets are analyzed thoroughly to suggest what the adversary's future weapons might be.

Yet, even complete weapons data could not anticipate the unpredictable "human" factors. Military actions may arise from unexpected sources, and random events may lead to growing armed conflict in many places. The combat may recede or recur at any time. Moreover, the process is partly circular: U.S. military preparations influence Soviet actions, which, in turn, influence U.S. actions.

Therefore, economists are not able to provide conclusive measures of the "right" level of military spending. But they have been able to clarify the most disturbing weapons issue of all: the dynamic process that causes the international arms race.

Dynamics of the arms race The arms race is a competitive process leading to excessive production and deployment of weapons. It occurs for the following reasons, largely related to the fact that the U.S.-Soviet relations are a form of oligopoly.

Because U.S. and Soviet policies are partly interdependent, one cannot define the efficient level of U.S. armaments without considering Soviet actions. The United States cannot realistically expect to attain perfect security, for such a high level of U.S. armaments would threaten the Soviets. The Soviets would react either by expanding their armaments to restore equality, or by taking some desperate action to remove U.S. superiority. The same logic holds for the United States.

Therefore, the two superpowers are locked in a need to reach approximate parity. Neither can seek clear superiority without causing the other to redress the balance. Exactly the same logic applies to any pair of adversary countries (such as Pakistan and India), and to groupings of allies. The result has been a continuing arms race, leading to over 50,000 nuclear weapons, the vast yearly spending on

weapons, and the large international weapons trade.

Concerning this process, economists offer three lessons: *First*, efforts by one superpower to gain superiority over its rival are irrational and ultimately dangerous to all sides. *Second*, a stable condition of peace requires both some guarantee against surprise attack and some assurance that any nuclear action will cause severe damage to all. But *third*, an absolute guarantee against nuclear actions can also be costly. By neutralizing the risk of nuclear war, it makes conventional warfare, without fear of escalation, more practicable. The nuclear threat itself is horrific, economists note, but it has prevented direct warfare between the two superpowers. That has probably prevented large amounts of conventional war costs and destruction that might otherwise have occurred.

Yet, the arms race continues, with its large opportunity costs. The economic burdens increase the severity of other world economic problems. Moreover, those armaments are likely to be used sooner or later, causing even more loss. Even limited warfare can severely disrupt the economic systems of both adversaries. Indeed, just the risks of future war can discourage productive long-term investments.

The economic basis for a volunteer army

A recurring urgent issue of public policy is whether young people should be subject to a military draft in peacetime. One doctrinaire position is that everyone has a patriotic duty to serve the country. Another is that drafting people is immoral, since forced military service is a form of involuntary servitude.

The most common argument on behalf of the draft in preference to a volunteer army is that the draft is cheaper than the alternative ways of staffing the armed forces. Because draftees cannot decline to serve, their wages are much lower than those that would be needed to raise a volunteer army of the same size and quality. Proponents of the draft frequently argue that the government cannot afford a volunteer army. Generals often argue this way (although the general staff itself is all volunteer!).

Do you find this argument persuasive? If you do, read on, and see if your views are influenced by economic analysis.

Why don't you want to be drafted, even though being a peacetime soldier is not a hazardous occupation? Presumably, you think that you have better things to do. The cost to you of being drafted is the value to you of the next best thing you could do with your time; that is, your *opportunity cost*. If you are in college, your time must be quite valuable because you are paying a large price to be there, in terms of tuition and the earnings you are foregoing. Those neither in college nor in the military are in civilian jobs, earning rates of pay that correspond to the value of their productive work. If they are taken from jobs that pay $15,000 to serve in the army at $6,000 per year, the true cost of their time is still $15,000.

An economist would calculate the true cost of an army of drafted soldiers not according to the dollars that they are actually paid in wages (such as $6,000 apiece), but according to what the same draftees would have to be paid to attract them into the army voluntarily. This value based on a full payment measures the opportunity cost of their time. Of course, this calculation attaches a far higher cost figure to the draft army than the actual dollars paid by the army to its draftees. But since draftees are forced to serve, their pay has little to do with the economic value of their time.

The draft does not save money. It simply reassigns the true costs of military ser-

vice, away from taxpayers and onto the conscripted youths. The youth who could earn $15,000 per year is taxed $9,000 per year by the draft.

That the draft subtracts from civilian production was recognized over 200 years ago by Benjamin Franklin. In opposing the impressment (the forcible drafting) of sailors, he stated the economic principle clearly and even estimated the size of the hidden tax:

> But if, as I suppose is often the case, the sailor who is pressed and obliged to serve for the defence of this trade at the rate of 25s. a month, could have £3. 15s., in the merchant's service, you take from him 50s. a month; and if you have 100,000 in your service, you rob the honest part of society and their poor families of £250,000. per month, or three millions a year, and at the same time oblige them to hazard their lives in fighting for the defence of your trade.

To show the economic effects precisely, we use the standard analysis of supply and demand for labor in the armed forces. Like any other line of work, the military services have a supply of potential workers, as illustrated in Figure 9, using hypothetical numbers. The number volunteering for military work depends on the rates of pay offered. There is also a demand curve, reflecting the value of the work done. The curve slopes down in line with the law of diminishing returns. The 500,000th soldier adds more value to natural defense than does the 3 millionth or 5 millionth soldier.

The equilibrium at Point A provides the correct number of soldiers. The marginal soldiers contribute a value (demand) which just equals the cost of paying them (supply). Beyond that level, the value of additional soldiers would be less than

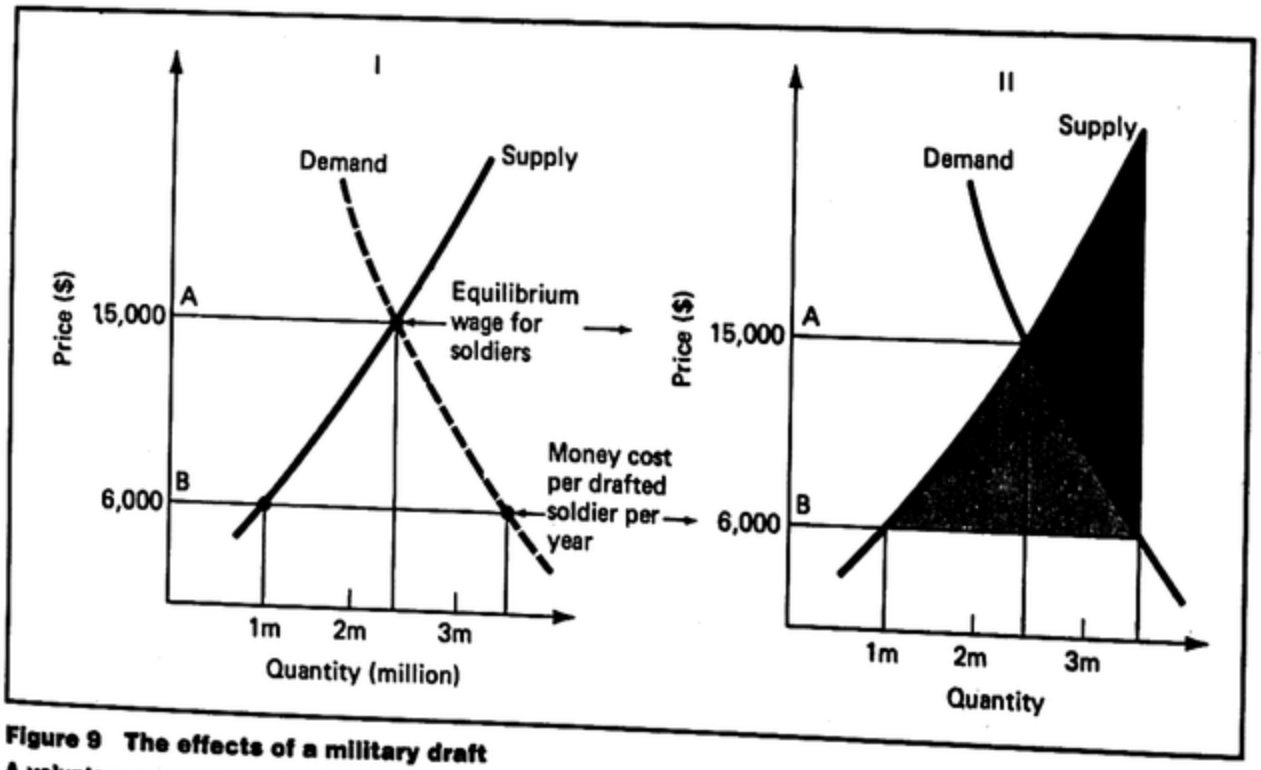


Figure 9 The effects of a military draft

A volunteer army succeeds when there is an equilibrium of supply and demand, at a cost per soldier of $15,000 per year. If there is a draft, the cost per soldier appears to be only $6,000 per year (covering room and board and a small rate of pay). But the real cost to society also includes the net loss of production value, if the drafted soldiers had been in civilian jobs. Area 1 shows that loss: it is subtracted from the national output. The drafted soldiers lose even more heavily. Their loss of income is all of the shaded areas, numbered 1, 2, and 3.

their cost. At Point A, the pay rate is high enough to sustain a volunteer military of 2.5 million total workers, earning $15,000 on average. Workers make free choices between civilian and military jobs. Thus, opportunity costs guide the decisions, and the nation's total production is maximized.

The effects of a military draft are readily contrasted in Figure 9. Military expenses are kept low, at level B, just $6,000 per year (covering costs plus a small wage). Only 1 million workers volunteer to serve at that low rate of reward. Meanwhile, it appears to the military management to be worth hiring 3.5 million soldiers at that low price. The resulting "shortage" of 2.5 million soldiers seems to make some kind of military draft necessary. But the draft would be unnecessary if the equilibrium price of A were paid.

Note how this diagram shows clearly the major points made. The suppression of military pay to B causes military officers to want a larger army: The value contributed to the military by the soldiers (shown by the demand curve) is above $6,000 in the range between 2.5 and 3.5 million soldiers. Therefore the use of those soldiers (at a *money* cost of $6,000) appears to be justified. But the drafted soldiers (between the 1 million volunteers and the 3.5 million total) would otherwise have held civilian jobs; the value they would have produced in those jobs is shown by the supply curve, which lies far above the $6,000 level of money costs for drafted soldiers.

The economic losses caused by the draft are shown in Panel II. The drafted soldiers lose all of shaded areas 2 and 3. They would have been paid their value in civilian jobs, which is shown by the supply curve. Instead, they are only paid $6,000. The entire economy loses the civilian production of those workers between A and B. It does gain the added military value shown by the demand curve between A and B. But the loss is more than the gain. The shaded area 1 shows that net loss.

In short, supply-demand diagrams show more precisely the relative gains and losses from the draft. They can also illustrate the effects of various elasticities of the curves. For example, if you redrew Panel II with supply much less elastic and demand more elastic, the draft's economic cost would become much larger.

Only during a war, when the military needs preempt civilian ones, does a military draft fit the criteria for economic efficiency. But that is partly because the situation is unusual and out of equilibrium. In normal peacetime conditions, where markets and free choices function reasonably well, a military draft is economically inefficient.

Since 1972, the army has been staffed on a volunteer basis, in line with this economic logic. But during 1978–1982, there were calls to revert to a draft, and steps to reinstate it were taken in 1980, at the urging of military officials seeking to reduce their budgetary requirements for military pay. Those favoring a draft note two defects of the volunteer forces. ***First,*** **the volunteer recruits are drawn heavily from the ranks of the poor.** ***Second,*** **there is a high rate of turnover in many of the more skilled military jobs.** Many young recruits sign on for the bonuses and pay, obtain their training, and then go to better-paying civilian jobs instead of reenlisting. For them, the military is a vocational school, not a career. That raises the services' training costs, while leaving them with fewer reliable, long-term professional members in positions requiring special skills.

Therefore, the "deterioration" of military personnel—combined with budget pressures and rising tensions over Afghanistan, the Middle East, and Central America—led to a reversion toward the draft.

Reinforcing it was a wish to "prove our resolve" by showing our "willingness to sacrifice."

Economists do not deny the *logic* of these points, but they insist that in these *matters of degree*, the opportunity costs of the draft are large, and proper economic incentives (i.e., improved levels and design of payments to soldiers) could solve the problem. Any desired size and mix of personnel can be achieved by making the economic incentives sufficiently high.

Consider the two asserted problems: low skills among ordinary soldiers, and an excessive rate of turnover in skilled positions. Economic research has shown that the supply of labor for these positions is price elastic: Higher wages will induce a higher quantity of workers. The elasticities are large enough to indicate that higher salaries have a strong degree of drawing power. Indeed, because of those elasticities, the present low military wages are deterring volunteer soldiers.

The problem arises in the archaic military pay system. Rather than being paid for the type of work they do, members of the armed forces are paid in line with their rank and length of service. Those holding the same rank and seniority are usually paid at the same rate, no matter how skilled their job.

This contrasts radically with the wage patterns prevailing in the private economy. There, pay is governed mainly by the marginal revenue product. Pay is aligned with productivity (at least approximately), in line with skills and scarcity. As the relative scarcity of each category of workers changes, wages adjust accordingly. Therefore, wages largely fit and respond to opportunity costs.

This is not true in the military pay system, in which all personnel can rise through a standard pay scale (grade E-1 up to E-9) regardless of the nature of their work. Many soldiers in low-skill jobs are now overpaid for the value of their work; many skilled workers—trained at great expense to the military service—are underpaid and therefore leave.

The needed reform is to set different pay grades in line with skill levels. Thus, supply clerks might be in grades E-1 to E-6, while radar technicians might be in grades E-3 to E-9. There would still be equal opportunity for promotion, but differences of pay would reflect skills.

The pay differences would be fitted to private-market wage differentials. Reenlistment bonuses could also be designed to induce soldiers with scarce skills to remain. By correctly designed incentives that reflected true economic scarcities, the military could achieve any desired mix of skills and other attributes.

Concurrently, military pensions could be revised. The current system encourages all soldiers to serve 20 years and then retire at a substantial pension. An efficient system would selectively encourage the scarcer workers to remain and retire at the ages common in the economy. Such a system could provide fully for old-age security, while encouraging the needed variety of skills, and at a much lower level of total cost.

Altogether, an efficient, selective pay system could make the voluntary approach work, probably at a *lower* total budgetary cost than in the present basis. ***The need is simply to align pay rates with the basic economic conditions of marginal productivity, scarcity, and opportunity cost.*** A draft superimposed on the rigid pay system would simply compound the economic inefficiency and eliminate free choice on a large scale.

Summary

1. These cases illustrate the main issues in public finance.

2. Benefits need estimating and costs need measuring. Then efficient choices can try to reach the efficient margins, even if only approximately.
3. Those simple criteria remain valid, but special care is also needed in designing the inner nature of the programs and the kinds of economic incentives they apply.
4. Commonly, the treatments also involve issues of competition and monopoly.
5. Moving toward a more competitive market basis (as in schools, pollution, and military hiring) can often clarify choices and activate the right incentives.
6. Even the simple analysis in this chapter has shown that some inefficient public programs have arisen and persist. Several of them are on a grand scale, and their wastes run into many billions of dollars each year.

Key concepts

Social benefits
Social costs
Cost-benefit analysis
Economic incentives

Questions for review

1. a. Describe the *voucher* system for education.
   b. What are some of its costs and benefits?
2. Describe one possible method by which pollution by private firms can be regulated. Will this method balance the marginal costs and the marginal benefits involved in controlling pollution? Explain.
3. a. How do the wage patterns in the military differ from those in the private sector?
   b. How does this difference tend to create inefficiency in allocating military labor?

21

Natural Resources: Concepts and Policies

As you read and study this chapter, you will learn:

- the economic concepts of natural resources and conservation
- how to define the optimal rate for using natural resources
- conditions under which private markets may optimize resource use
- the main issues in agricultural economics and farm policies
- how prices and elasticities can avert future resource crises

You have undoubtedly seen photographs of Earth taken by the lunar astronauts. There is Earth, a sphere glowing blue and white and floating in the dark void of space. It is a beautiful, haunting sight, which no human being had ever seen before the 1960s.

You may also have noticed how small and lonely Earth looks. Until 1969, humans had always been on or near Earth's surface. There, the planet seems enormous, with its endless shining seas and vast plains. Now, Earth can be seen and understood for what it is: "Spaceship Earth" holding limited amounts of resources and room. Some of its precious resources are, in fact, being used up by the growing demands of industrial growth: topsoil, water supplies, metal ores, coal, sulfur and other minerals, gas and oil—especially oil. Acid rain as strong as vinegar now falls on forests and lakes, and toxic wastes are reducing the purity of many rivers. The depletion of natural resources has become a leading economic problem, especially in the United

States. When the resources are gone, what then? Economic doomsdays have been forecast for the next century, even as early as the year 2020, fewer than 40 years away.

This resource problem is one that economists have long studied. Adam Smith, Ricardo, Malthus, and other early economists were deeply concerned about the inevitable depletion of resources by economic growth. In fact, that is how economics came to be dubbed the "dismal science": It contemplated a dismal future of growing scarcities under the pressure of rising population and production. In the mid-to-late 1800s, however, industrial growth boomed, science produced many highly productive inventions, and vast pools of oil and lodes of minerals were discovered. Science seemed able to banish scarcity. Economic growth might well go on forever, it seemed. The "conservation movement" struggled against a prevailing industrial optimism.

Events since the mid-1960s have dispelled much of that optimism, and the adequacy of natural resources is now widely regarded as an economic threat to future growth and social stability in the world. This chapter presents the fundamental economic concepts, expanded from the brief discussion of conservation in earlier chapters.

We first present the tools for determining the optimal rate of using exhaustible resources. Using these concepts of conservation, we then take up a specific problem: America's farm policy. Finally, we discuss the likely future course of energy prices.

Basic concepts

Natural resources come in many kinds, as Table 1 shows. Some are abundant, like sunlight and seawater. Others are renewable and can be efficiently harvested virtually forever: forests and farm crops, for instance. Still other resources, such as oil, ores, and coal, are strictly exhaustible. Once used, they are gone forever.

Table 1 The main types of natural resources

Nonrenewable
Fuels (coal, oil, gas), land, ores, chemical deposits
Replaceable at great cost
Soil, wilderness, certain rivers and lakes
Renewable
Other rivers and lakes; urban fresh air
Self-renewing
Forests, fisheries, other "crops"
Virtually inexhaustible
Rural fresh air, solar energy

Should the exhaustible resources be saved, rather than used? If they should be used, how rapidly? Such issues appear to be complicated, involving many special features. No single best rate of use applies to all kinds of natural resources. Each one needs careful study.

Yet, two major concepts are fundamental to them all: economic rent and conservation. **Economic rent is a payment made to an input that would be supplied even if no rent were paid.** Most natural resources are paid an element of economic rent, as growing demand causes their prices to rise. ***Conservation*** is the other main concept involving natural resources. We now explain its meaning and uses.



Conservation: Reaching the optimum rate of use

The term "optimum" here implies *social* efficiency: the best net use of resources for society as a whole, over the relevant span of time. The general meaning of allocative efficiency was presented in Chapter 16. Essentially, it requires that all resources be used in each time period up to the point where their marginal benefits just equal their marginal costs. Natural resources

provide a special case within this general rule.

However, costs and benefits must be very carefully defined when they are applied to natural resources of limited amounts. The economic aim is to use efficiently—and equitably, both *within* each generation of people and *among* generations as the decades and centuries pass—each physically limited, depletable resource. Efficient use will yield the maximum net value for the resource over time, according to several factors we will discuss. Equity involves difficult problems because the resources used up by present generations are denied to future generations. Since the future inhabitants of the world are not here now to urge that resources be saved for their use, present generations may selfishly consume these limited resources *too rapidly*.

Conservation does not equal preservation Yet, the hoarding of resources for future use can err on the other side, toward *too slow* a rate of use. The goal is to strike the right balance between present and future. To define this goal, you can best begin by recognizing that each resource is an asset, a physical stock, that may have economic value. It can be held in its present form or used at some rate, either for present consumption or for investment.

Decisions about the use of resources are basically judgments about their future worth, either as left in their natural state or as converted to some other form. Physical *preservation* is only one alternative among the ways to conserve a resource. The efficient use of a natural resource often requires that it be used up. The economic task is to define the efficient rate of usage.

One common fallacy is to regard the present resource base as a fixed inventory that, once exhausted, will leave society with no means of survival. A related fallacy is that physical waste equals economic waste: that it is wasteful to use materials in ways that make them disappear. This attitude can lead to devoting $10 worth of work to "saving" a few cents' worth of paper, scrap metal, or bottles. ***Neither hoarding nor physical recovery is synonymous with conservation.***

Consider a decision on the use of iron ore. It can be kept in the ground for future use, or it can be mined, smelted, and fabricated to create new productive machinery *now*, with a consequent expansion of economic capacity to produce. Though it is physically destroyed or altered in creating the machinery, the ore's economic value is enhanced. Some rate of current use for that purpose is clearly justified.

The *logic* is clear; only the matter of *degree* is to be determined. The economist's task is to distinguish such productive uses of resources from those that deplete them "too fast." How can that ***optimum rate of use*** be determined?

The optimum rate of use can be defined precisely as a matter of logic, even though it is often difficult in practice to determine the exact values for each resource. ***The optimum rate is the speed of usage that will maximize the net present value of all future uses of the resource.*** Consider the case of an ore that must be mined and processed. Its net present value is found by the following steps:

1. Predict the *physical amounts of usage over future years*, under what you think is the best set of methods for mining and processing the ore. As illustrated in Panel II of Figure 1, the probable best approach would be to use the resource at a constant rate until the year 2100, when all of the known reserves of it will be gone
2. Predict the *prices at which the resource will sell.* In Panel II of Figure 1, a rising trend of prices is predicted.

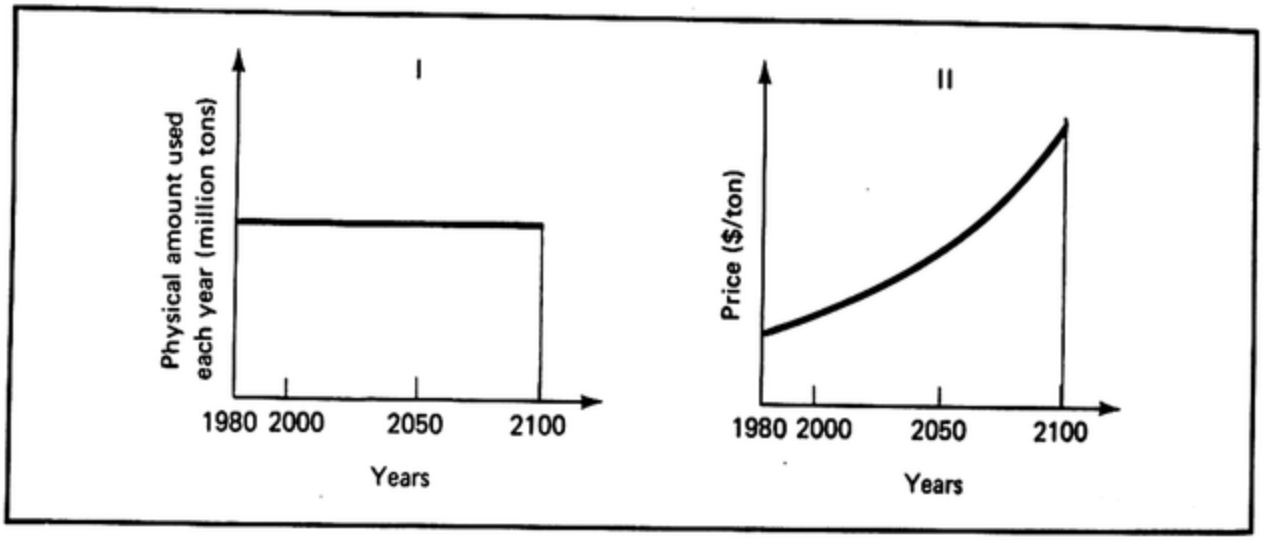


Figure 1 Two steps in choosing the optimum rate of using a resource

3. Multiply the prices times the physical volumes to obtain *the flow of future values from using the resource.* They are illustrated in Panel I of Figure 2. But these flows are not the final basis for setting optimum resource use because they ignore the element of time.
4. Therefore, you must discount the future values by some reasonable rate of time discount. This will probably be an interest rate in the range of 5 to 20 percent, although the best rate of discount may be highly debatable, as we will soon see. Panel II of Figure 2 illustrates the effect of using some reasonable rate of discount. The result is a single figure summing up the total present value of the total future time-discounted values from using this resource. That is given by the area under the solid line in Panel II of Figure 2. But even this is not the final answer. You must also allow for the costs of obtaining and using the resource.

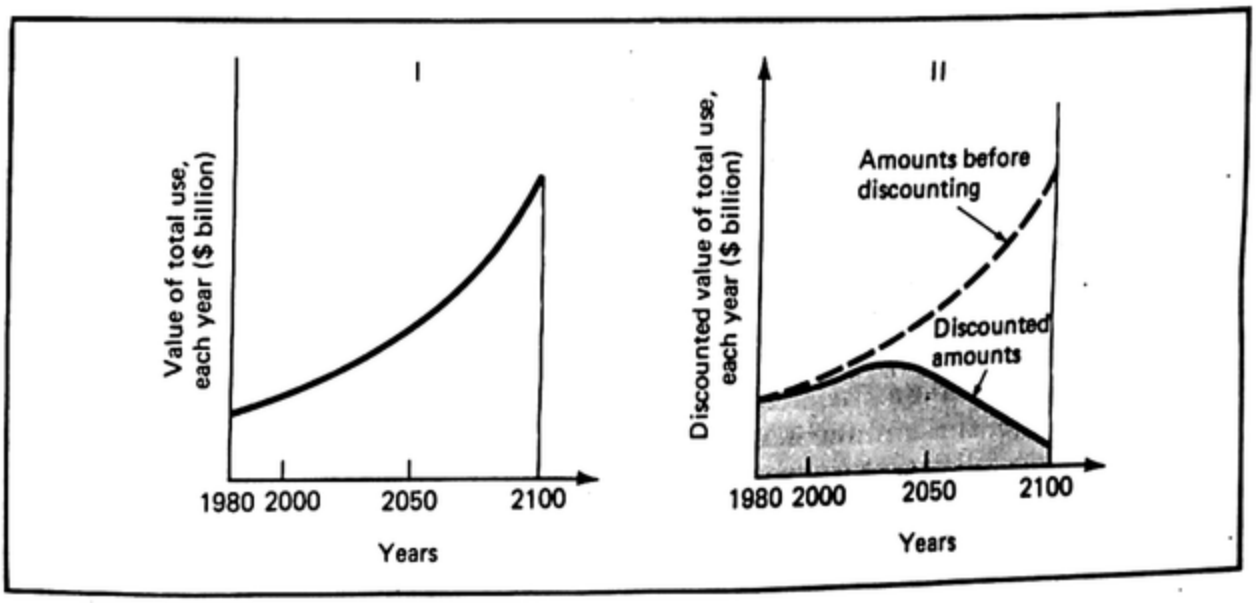


Figure 2 Calculating the discounted value of future use of the resource

5. Thus, you must next *estimate the future costs of extracting and preparing the resource* between now and the year 2100. That involves guesses about the increasing physical difficulty of mining the ever-depleting resource in deeper and deeper veins. You must also guess how far the future progress in mining methods may offset some or all of that increasing physical difficulty. Perhaps new lasers or conversion techniques will sharply reduce the costs. Or, instead, rising fuel costs may sharply raise the costs of mining and processing the resource. Panel I of Figure 3 shows an estimate of these future costs.
6. Then you must *discount the future costs for time,* just as you discounted the values of the resource flows. You will, of course, use the same discount rate, whatever it is, on both the costs and benefits. The result is illustrated in Panel II of Figure 3.
7. Now you can *subtract the discounted costs from the values* of the resources, to obtain the ***discounted net present value*** of the resource from that plan. That step is also illustrated in Panel II of Figure 3. Let us suppose that the resulting figure is $500 billion. That measure of the net present value is precise, even though many of the elements going with it —the reserves, the future prices and costs, the "correct" discount rate, and so on—were guesses. Altering those values would change the final sum. Yet, since those estimates are the best you can do, the final figure is significant.
8. But there is still one more step: *You must compare this value with* alternative *reasonable plans for using the resource,* to be sure that this first plan really does give the highest net present value. Thus, you try various alternatives, as in Panel II of Figure 4. Plan *B* has a longer, slower period of use, lasting until the year 2200. Plan *C* uses it all by the year 2050. Plan *D* starts at a slow rate but then accelerates. Plan *E* starts fast but tapers off.

The net present values of each plan are shown in Panel II of Figure 4, using the same price, cost, and discount assumptions (for simplicity, we have omitted the intermediate calculations). Suppose that the net discounted present values for each plan are: Plan *B* $600 billion, Plan *C* $460

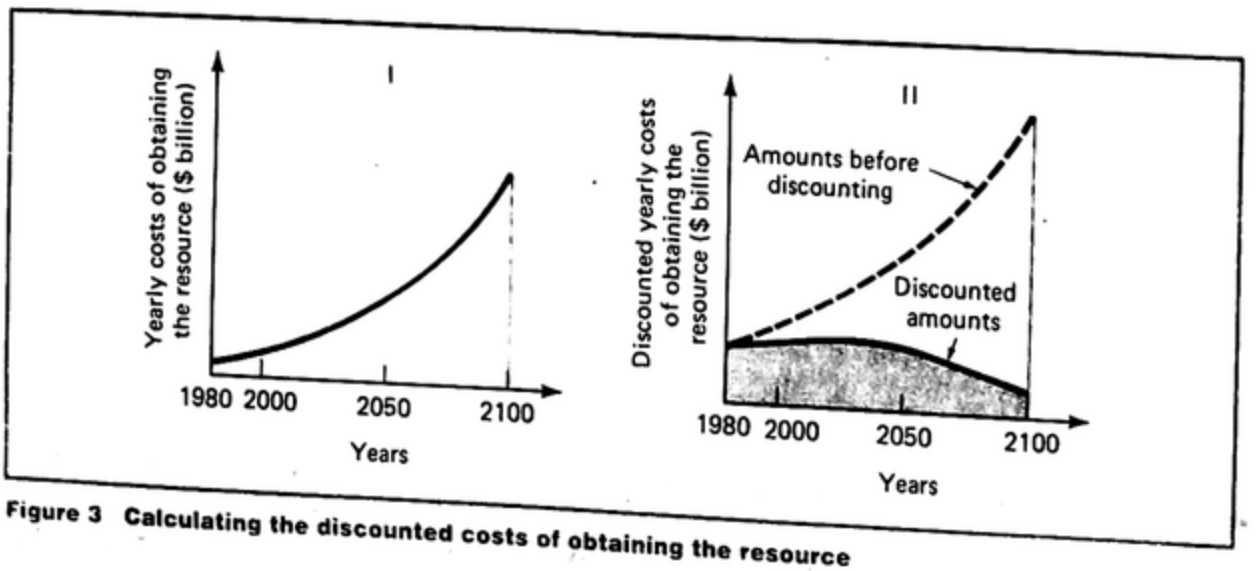


Figure 3 Calculating the discounted costs of obtaining the resource

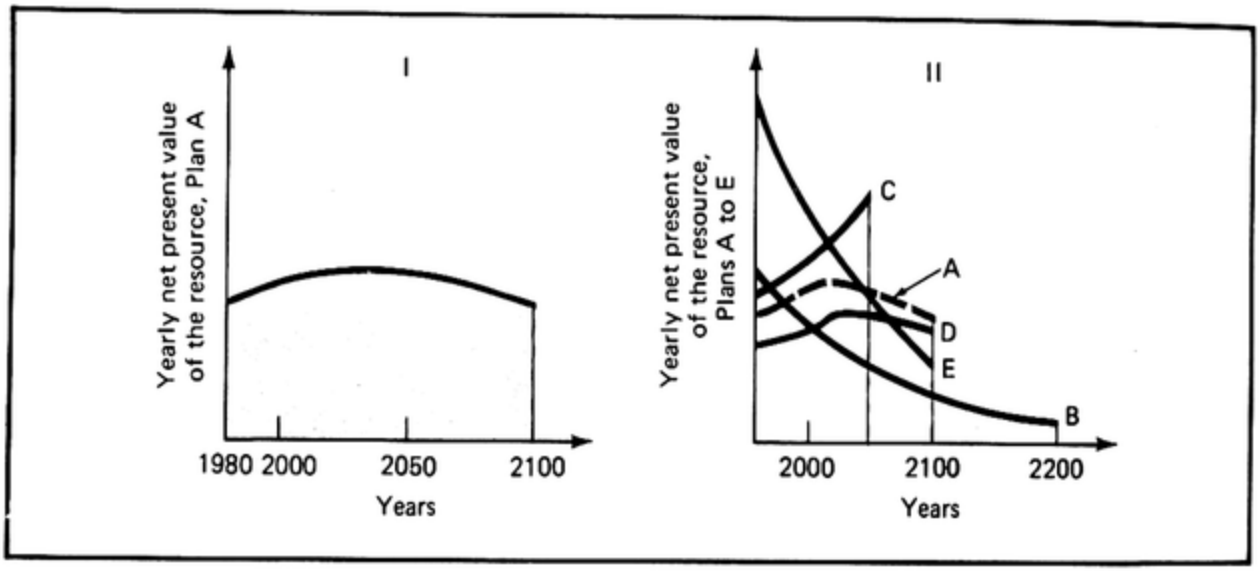


Figure 4 Comparing final present values for alternative uses of the resource

billion, Plan *D* $480 billion, and Plan *E* $680 billion. Evidently Plan *A* was not the optimum, for its $500 billion is exceeded by both Plan *E*'s $680 billion and Plan *B*'s $600 billion. Plan *E* is the optimum rate, at least until further calculations reveal a better one.

Note that Plan *E* uses the resource rapidly now, before the time discounting of distant years begins to weigh heavily. Note, too, that the process is similar to the profit-maximizing calculations made in the appendix to Chapter 7. Whenever future plans are compared, an economist calls for time discounting of future values.

Finally, note that the optimum rate of resource use merely compares costs and benefits. Here, there were some complex elements and a long time horizon. But the basic reliable concept is the same as always: Define costs and benefits correctly and then compare them.

Five determinants of the optimum rate

Now consider which parts of this process are most decisive. There are five main elements determining the optimum rate for each resource: current prices compared with future expected prices; the rate of interest; the costs of finding, using, and renewing the resource; predictions about future technological change; and the ethical weights used to compare the needs of present and future generations.

1. **Present and future prices** Prices directly indicate economic scarcity and, thus, relative value. If a resource's price is expected to rise sharply above its present level, that tilts the choice toward holding the resource for its later, more valuable use. Conversely, if a resource's price is expected to drop steeply, then it might as well be used more copiously now. The falling expected price indicates that its future value will be small. Therefore, saving it is not economically valuable. The expected future price trends are, of course, only estimates, which require judgment. These are not official sources or infallible predictions.

2. **Time discounting and the rate of interest** When people discount the future sharply compared to the present, they are assuming high interest rates. That encour-

ages using resources faster now, rather than holding them for future use. Comparisons like those in Figures 1–4 are usually highly sensitive to alternative interest rates. Thus, a higher rate of interest would have made Plan *C* relatively more attractive, for it would have cut the later discounted value of the other plans more sharply.

Consider another example: A fine stand of walnut trees may promise to sell for $700,000 in 1999, compared to just $200,000 in 1983. At an interest rate of 10 percent, the 1999 sum has a present value in 1983 of $152,340; the trees should be kept standing until 1999. But if the interest rate rose to 15 percent, the present value in 1983 of the 1999 sum of $700,000 would fall to $74,805; then, the trees should be harvested without delay.

In general, higher interest rates favor a faster present use of resources. Low interest rates favor a slower rate of use, over a longer time span.

3. Direct costs of finding and harvesting the resource Resources do not rise spontaneously out of the ground, smelt themselves, or carry themselves to market. There are usually costs of discovering, mining, converting, planting, harvesting, or transporting them. These direct costs often influence how rapidly the resource is exploited.

For example, the search for oil and gas proceeds just up to the margin at which the value of finding them is balanced by the cost of the search. Fishers take their ships out only so long and so far as the marginal cost of catching the fish equals the marginal value of what they can catch. Farmers use their soil only to the margin where the extra crop's value justifies the cost of the extra resources used. American settlers pushed out into virgin land only so far, in each decade, as the marginal value of the land justified the costs of clearing and planting it.

This general principle is clear and powerful, applying to all resources and all periods. The analysis takes a very specific form in dealing with resources that are cropped, such as fish, trees, and farm crops. *The optimum yield in each period is a balance between the sales value of the crop and the costs of raising and harvesting it.*

In Figure 5, the rounded curve shows how crop yields respond to increasing cultivation. Variable inputs are grouped together on the horizontal axis. Although they provide a rising total crop yield, the familiar diminishing marginal returns set in immediately. Eventually, at Point *B*, they grow strong enough to cause even total returns to fall. If this were fishing for

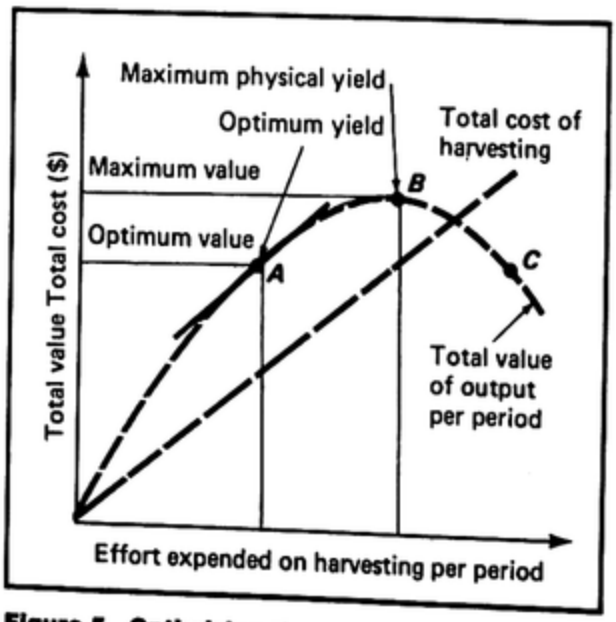


Figure 5 Optimizing the use of a renewable resource

By harvesting more intensively, the fishers can catch more fish—but only up to Point *B*. Beyond that level, they reduce the total catch by overfishing.

But *B* is not the economically efficient level. That is reached earlier at Point *A*, where the marginal value of extra fish just equals the marginal cost of catching them. Therefore, the efficient level of fishing effort stops short of that necessary for the maximum physical crop of fish. This principle applies equally to other harvested resources, such as grain and lumber.

cod, for example, Point *B* would show the maximum physical catch. Heavier fishing (to the right of Point *B*) would deplete the schools of fish and reduce the total catch, as at Point *C*.

The optimum level of harvesting this renewable resource is precisely at Point A, *where the marginal value of the extra yield of fish just equals the marginal cost of catching them.* (Remember, marginal cost is the slope of the total cost curve; marginal revenue is the slope of the total revenue curve. At Point *A*, the slopes of the total cost and revenue curves are equal.) Notice that *the optimum yield at* A *is below the maximum possible physical yield.*

Think of every crop in this fashion, comparing costs and revenues from harvesting efforts at the margin. In this light, "maximum yield" is not a desirable goal. The optimum level is less than that.

4. Future technological change The best present use of a resource usually depends on its present alternatives, which can change as time passes and technology advances. The future uses of the resource, therefore, depend on future conditions, which can often only be guessed at.

For example, oil's value has risen sharply because there are few good substitutes for it as a fuel for many machines. But fusion power might be developed to replace oil for many uses, such as generating electricity and fueling industrial ovens and boilers. Accordingly, oil would be less scarce and, therefore, less valuable in the future. The present value of oil, gas, or coal hinges on such predictions about future technology. Major technical advances can deflate the value of resources by making them less scarce.

Predicting technology changes, although notoriously difficult, is an essential part of making choices about resources. Often, estimates of future technology sharply conflict. Thus, some experts foresee breakthroughs in fusion and laser technology that will radically reduce the cost of energy from the next century on. Other experts expect the opposite: The breakthroughs will not occur; nuclear energy's problems will intensify; and energy scarcity will become increasingly severe.

Energy is only a dramatic instance of the element of gambling and uncertainty that occurs in all predictions of future technology, as in metals, foods, and forestry. The predictions are constantly being revised in the face of innovations and new information. But it is often virtually impossible to predict technology beyond 20 years into the future.

5. Ethical weights among generations How strongly should our present interests weigh, compared to those of our children and their successors? Each generation inherits certain resources, uses some of them, and passes the remainder on to the next generation. The present generation is no exception: It is shrinking the resources available to future generations though innovations are also opening up other opportunities. What is a fair balancing of ***intergeneration choices,*** between our interests and those of our numberless unborn successors?

According to careful estimates, in the year 2100, there will perhaps be 11 billion people, compared to 4.5 billion now. This larger population will need resources even more urgently than we do. Yet, most of the oil and gas will be gone, farmland will be further eroded or put to urban uses, and the stocks of ores and many other resources will be much smaller than they are now. The extent to which current use is irresponsible or excessive—or, possibly, slower than optimal—depends on the eth-

TEXT BOOK LIBRARY
Acc No. 15358
SRINAGAR

ical weights used to compare our needs with those of the people who will be alive then.

There is no definitive way to formulate or apply those weights. Roughly speaking, *giving a higher weighting to the interests of future generations means using a lower rate of time discount.* At the extreme, we might weight future people's needs so much more heavily than our own that a *negative* rate of interest would be suitable in assessing the present values of resources. Then the present value of using 100 billion barrels of oil in 2001 or 2050 or 3050 would exceed the value of using that oil now.

More normally, a *positive* rate of discount applies, but it may be high or low. If we apply a high rate, we will use up resources relatively rapidly, leaving less of them for future generations. They may then curse us for our extravagance. But, being long dead, we will not have to suffer for having served our preferences rather than theirs. Indeed, we will have had it easy. That is the problem in attaining an optimum rate of use among generations: Each generation can ignore posterity with impunity.

Private markets can optimize the use of resources, except. . .

Now return to the general case of a resource that is privately owned. Economists offer a clear, optimistic lesson about the conservation of such resources: ***Private markets operating with reasonably complete knowledge and rationality can meet the social criteria for conserving resources over time.*** The owners will be guided both by their profit-maximizing motivations and by the objective conditions prevailing in financial and industrial markets. These will tend to reflect precisely those social valuations of time preference, productivity, and expected innovation that determine the optimum usage rate. Moreover, it is in the interests of the resource owners to seek out accurate information on these magnitudes and to apply them in their own decisions. Therefore, the private market disposal of natural resources can optimize their use.

This result holds only for competitive markets. Monopolists will usually hold the rate of resource use *lower*, by restricting output and raising price in the present. That will lean toward too-slow use, rather than toward using resources too quickly. Yet, because even monopolists will want to maximize the value of resource use in the long run, the restrictive effect may distort their choices only slightly.

Altogether, then, when there are rational choices in private markets, the prices of natural resources will tend to anticipate future scarcities. The logic applies to all resources: land, oil, minerals, coal, gas, and the rest. Their market value moves with changes in the expected stream of future rents, which, in turn, reflect the expected shifts in demand and supply. Therefore, if people's expectations about future prices should rise, they will quickly bid up the capital value of the resource. Prices rise in anticipation of coming shortages. That price rise acts, in turn, to reduce the usage of the resource, if demand is at all elastic.

Therefore, the rising scarcity leads to (1) a rise in price ahead of the actual shortage; (2) possibly a cutback in the quantities used; and (3) windfall capital gains for the resource's owners. Such "forward pricing" acts to smooth the onset of rising scarcity: The rising prices reflect and enforce the present need to conserve the resource more stringently.

To this degree, the owners and users of resources will act in accord with the genuine social costs of their use of resources

(unless there are externalities. Further, as unexpected changes occur in these predicted future scarcities, the prices of resources will adjust quickly and automatically. Therefore, both the present and future scarcities, even if they are changing and uncertain at each point, may be reflected as fully as possible in the prevailing prices and rates of usage. This will occur spontaneously, without conscious or detailed social planning. In short, economists conclude: The "Invisible Hand" extends to conservation.

Limitations and biases However, this optimum result depends on strict conditions, which may not be met. Six such cases follow:

1. COMMON-PROPERTY RESOURCES Some resources are not individually owned. Thus, no price for using them can be levied by a specific owner. As a result, they are available at a zero price to whoever captures them. That can lead to competitive overuse and destruction, as each user maximizes its own profits by taking the resource rapidly. Fish are one example. Oil and gas in an oil field that can be tapped by different landowners who own plots of land above the oil field are another. Each user is oblivious of the conditions shown in Figure 5, which reflects the total use of the resource. Since the competitive user considers only its own interests and is aware that the resource will dwindle, the incentive for rapid removal is increased.

The net effect is often a race to capture the resource. If the resource is harvestable, as fish are, the process will exceed the optimum rate and possibly reduce the total catch or even render the species extinct (at the right end of Figure 5). If the resource is a fixed stock, such as oil, the current rate of extraction will be raised well above the optimum. Moreover, the oil field itself will be harmed: As the separate owners drill many wells to get the oil out faster, the excessive numbers of wells will cause the underground pressure to fall, thereby raising the costs of extraction.

The corrective to this problem is to create a monopoly: to unify the control of each such resource in one owner, so that the optimum technical pattern and rate of usage can be designed and applied. Here, monopoly is clearly preferable to competition. Of course, society should also prevent the monopolist from causing the harms that we reviewed in Chapter 10.

2. DISCOUNTING AND MYOPIA The private rate of discount may be too high. As Figures 2–4 illustrated, this encourages a current rate of use above the optimal rate. The high rate of discount gives more weight to the present generation's interests than to those of future generations.

Actual rates are more likely to be above the correct social rate of discount than below it. Ultimately, perhaps, a negative rate of time preference should be applied to some intergeneration choices. If population and income levels continue to grow, and technology fails to provide new methods, then the pressure on resources may far exceed anything now imagined. Therefore, in the long run, it might be optimal for us to put more value on future than on present use. At any rate, the social rate of time preference may be lower, perhaps much lower, than the rate established by private choices in private markets.

3. INADEQUATE FORECASTING Present users may simply fail to foresee future developments. This may reflect insufficient research or an inability to discern future change. There may be close interactions among the uses of resources that are not presently apparent to the various users. And some users may simply be careless or

irresponsible in their judgments. In any case, the result would be resource use that deviates from the optimal patterns.

4. POLITICAL INFLUENCES Specific taxes and other incentives may encourage an overly rapid use of resources. In fact, the use of almost every resource is affected by artificial incentives. There are special tax provisions for the extraction of almost all natural resources such as oil and ores. The use of land surrounding cities has been intensified by special tax provisions that favor suburban developments as investment tax shelters. Farm policies affect the extent and intensity of cultivation. Maritime policies affect the exhaustion of oceanic fish resources. These incentives usually induce a more rapid use of resources, even to the point of extinction.

5. EXTERNAL EFFECTS There are important externalities in the uses of many resources, so that private users ignore major social costs of their actions. This affects both the rate of withdrawal of natural resources and the degree to which the common environment of air, water, and habitat is degraded. Recognition of such externalities would require a slower and altered use of resources. The failure to do so has led to environmental damage from land, air, and water pollution.

6. DISTRIBUTION Finally, private market decisions are based on the existing distributions of wealth and income. Since resource users vote with their dollars, market demand will more strongly reflect the interests and preferences of the wealthy. This may conflict with broader social criteria. Some impacts, such as congestion and pollution, may affect the poor more acutely than the rich. And the poor have a greater need for common recreational space and facilities. Thus, an "unfair" distribution can result in undesirable outcomes.

Agricultural economics

Agricultural economics emerged in the 1920s as the leading branch of applied economics: Agricultural economists actively measured real demand and supply curves, estimated economies of scale, and even showed a dynamic cycle in farm prices and outputs. But agricultural economics is more than a technical field. It must grapple with some of the most dramatic issues and dubious public policies that are found in any sector. To present them, we begin with the basic conditions of farming.

Basic conditions

Agriculture is the growing of crops and livestock. Crops are sown, weeded, and harvested. Livestock are bred, raised and fattened, and sold. All of this occurs on a thin layer of topsoil, varying from a yard or more deep in the richest Iowa land to a mere few inches or less on the western ranges. To understand agricultural economics, one must recognize several main features of the agricultural sector.

Variety The agricultural sector embraces a great variety of conditions. Figure 6 indicates the main kinds of production, each with its own conditions of demand and supply. There are sharp differences among, for example, western cattle grazing, Florida orange growing, Virginia tobacco farming, and potato growing in Maine and in Idaho. Each uses special combinations of inputs (soil, fertilizer, equipment, etc.). Each has its particular regions, climates, sensitivity to weather, and classes of consumers.

Rising productivity Despite its old-fashioned aura, the agricultural sector in the United States has had rapid progress, with productivity rising more than twice as fast as in the economy as a whole. This high

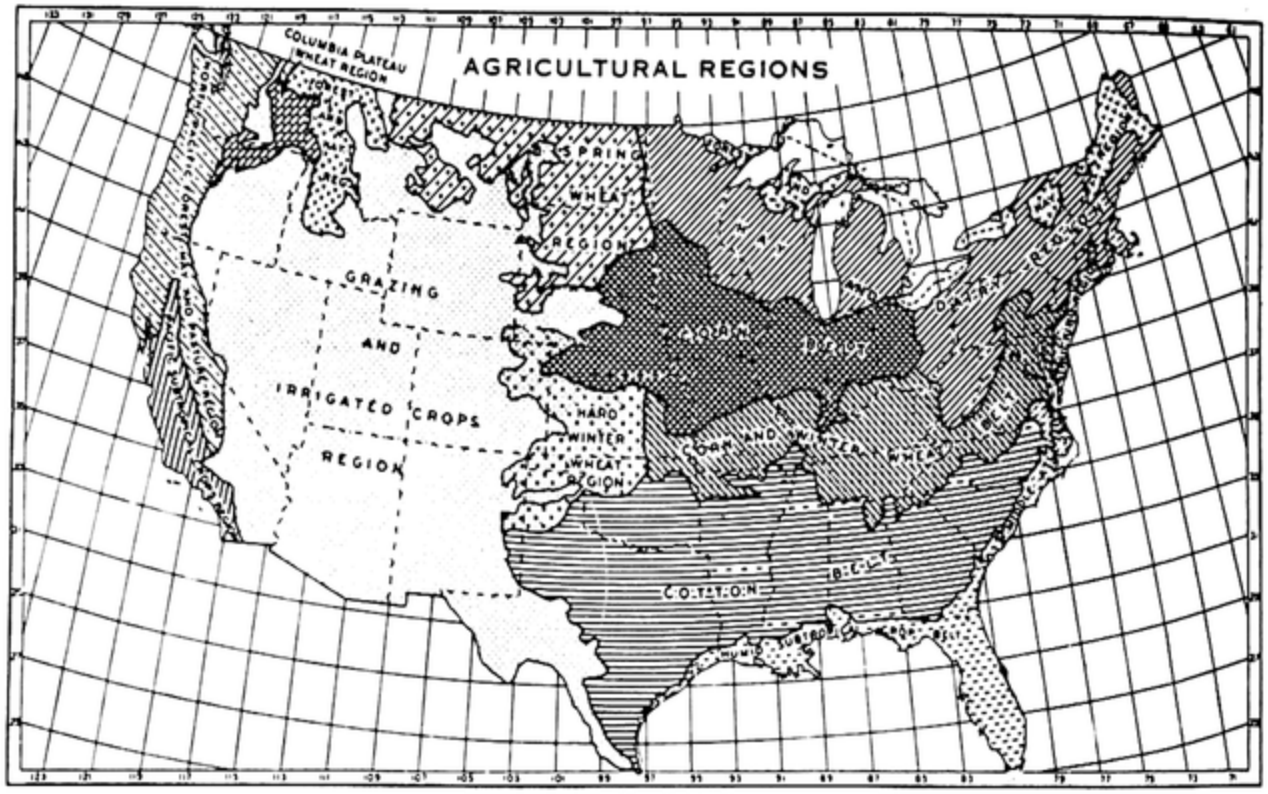


Figure 6 The main patterns of U.S. agriculture

Only the main types of crops and livestock are shown. Even so, it is evident that there is much variety within the agricultural sector.

productivity reflects rich soil and a favorable climate. But the advance of agricultural technology has also been a major cause.

The rising efficiency of agricultural technology reflects three main causes. First the improvement in capital inputs. Farm machinery has grown larger and more complex. Each unit of capital can produce a larger volume of output with about the same labor. Thus, a big combined harvester-thresher-baler can handle triple the volume that machinery of the 1940s could do, but still needs only one person to operate it. The equipment both raises efficiency and substitutes for labor.

The second cause is improved fertilizer, insecticides, and chemicals for livestock. In recent decades, much more and better fertilizer has been used, yielding higher crops and permitting more crop cycles during the year. Chemical insect and bacteria killers have also become widely used for ground and tree crops. Heavy doses of medicines and specially prepared chemical feeds are now routine for livestock.

The third main cause is better *breeding*. Since the 1920s, there have been major genetic improvements in crops that have given faster growth, hardier plants, better resistance to diseases, and larger yields per plant. Livestock breeding has also become more efficient and produced better animals.

Since these gains are continuing, agricultural productivity may continue to advance at unusual rates. Yet, there are two main limits to this advance: increasing energy scarcity, and the loss of topsoil and water.

Energy scarcities Farming has come to rely heavily on commercial energy sources. Heavy equipment requires gasoline and

oil; fertilizers and insecticides are made by energy-intensive processes. This trend was efficient when farm wages were rising faster than oil prices, for the changes substituted a cheaper factor (energy) for an increasingly expensive one (labor). But the sharp rise in energy prices since 1970 has reversed the comparison. Therefore, two sources of rising farm productivity (capital and chemicals) are being cut off by changed scarcity conditions. Simple relative prices, reflecting relative scarcities, are now moving farm technology away from energy-intensive methods.

Topsoil and water The loss of topsoil and water sources is also likely to limit future agricultural gains in productivity. Natural soil forms slowly by the continual growth and mingling of plants and bacteria. Human cultivation endangers that process by breaking the ground cover and applying chemicals. That, in turn, exposes the soil to erosion by wind and water.

Yet, by the best estimates, a century or so of cultivation has dissipated about one-half of the nation's topsoil. On average, the loss is at more than 6 tons of topsoil per acre per year. The total annual loss from all U.S. farms is approximately 6 billion tons, equivalent to peeling an inch off the state of Missouri.

The rate of loss has recently been increased by the heavy use of chemical fertilizers, weed-killers, and insecticides. In the short run, these chemicals have increased crop yields. But they undermine the natural fertility of the soil, replacing it with artificial chemical nutrients. The soil itself becomes sterile and grainy and is easily eroded. As a result, topsoil is lost both by physical erosion and by the depletion of its natural fertility.

Artificial fertilizing reaches diminishing returns in the long run, as the soil deteriorates. Yet, in the short run, farmers have been induced to increase their profits by using fertilizers heavily. Indeed, they have had little choice, since when some farmers adopted the method, all of them had to follow suit to remain competitive. The resulting farming methods reflect a short-range view of land values, using high rates of time discount.

Since 1973, the increased scarcity of energy has been making chemical fertilizers less economical, by raising their prices. Yet, much topsoil is now dependent on the continued use of those fertilizers. Therefore, the long-run role of chemical fertilizers in the use and conservation of land is in doubt.

At any rate, the remaining topsoil is vulnerable to substantial further erosion. The increasing scarcity of topsoil will, in turn, accentuate the future increase in farm prices.

Water poses a different problem, for it is usually not owned by specific users. It gathers in the underground aquifers that make up the "water table." As a common-property resource, it is open to competitive overuse by farms, factories, and towns. One user can sink a well, install a large pump, and drain the water supplies that are needed for many miles around. On a larger scale, huge irrigation projects in western states have already substantially lowered water tables and are raising the likelihood of permanent depletion and water shortages. The problem has grown acute in the Great Plains, in south-central Arizona, and in the San Joaquin Valley in California.

With the basic problems of soil and water in mind, we now turn to the specific economic issues that have troubled farming and inspired a variety of dubious government policies.

Farm policies

Farm incomes are often unstable from year to year, for two main reasons. One is

weather. Drought or hail can destroy a crop one year; good weather can give a bumper crop the next year. The second reason is the inelasticity of demand and supply for most farm products. As we saw in Chapter 5, shifts in those inelastic curves can cause marked changes in prices, which, in turn, cause farm incomes to fluctuate sharply.

Everyone agrees that, ideally, farm incomes should be more steady. The first programs to stabilize farm incomes were started in the 1930s, using the logical idea of a buffer stock or "ever-normal granary." Thousands of storage bins were built in the farming areas for use as reservoirs. The long-run equilibrium prices and quantities were estimated. In good years, the government bought and stored enough crops to keep prices up to the long-run level. In bad years, the bins were emptied of enough crops to make up the gap, thereby holding the price to the long-run equilibrium level. The cost of the bins and storage was small compared to the benefits of stabilizing farm incomes. The program worked well for grains, powdered milk, eggs, and several other crops.

But then the program was altered to serve another purpose: *raising and holding the price above the long-run equilibrium level.* The "price supports" raised farmers' incomes by increasing their total and net revenues. But they also hurt consumers by raising the long-run prices that they would have to pay for the foods, clothing, and other products made from farm outputs. By raising prices above equilibrium cost levels, the price support program caused inefficient allocation.

Besides being inefficient, the programs were also inequitable because they mainly benefited the farmers with the largest levels of output. The higher crop price is multiplied times the quantities sold. Since small farmers produce less, they also benefit less, even though they are the truly needy farmers. The benefits have flowed mainly to the big, already-prosperous farmers. Altogether, price supports have reflected political clout, rather than the use of clear, rational economic tools.*

This perversion of the original programs from *income stabilizing* to *price raising* has had several predictable economic effects. *First,* as Figure 7 illustrates, the higher price has led farmers to produce more over the long run and consumers to buy less. The resulting physical surpluses were enormous in the 1950s and 1960s. The Agriculture Department has had as much as $10 billion tied up in surplus stocks, at a yearly cost of $3 billion for storage, spoilage, and interest. After 1970, many of the stocks dwindled as farm prices rose and made price supports less necessary and less politically acceptable. Yet, the prices of several important farm outputs are still supported, including milk, cheese, eggs, tobacco, peanuts, sugar, and California oranges.

A *second* effect has been to raise the price of farmland. By making crops more valuable, price supports have made the land itself more valuable. That explains some of the surge in land prices, which has raised prices per acre more than 12-fold since 1950. These benefits are strictly a rise in economic *rent,* which farmers get for doing nothing more than their normal work. That rent is shown by the shaded trapezoid in Figure 7. Indeed, one could say that the benefits have gone to the land or to land ownership, not to farmers as such. In any event, these benefits, too, have gone mainly to large-scale farmers, most of whom were already prosperous.

Since the surpluses were a costly and embarrassing defect of the price-support

*For a thorough treatment of these issues, see Geoffrey S. Shepherd and Gene Futrell, *Marketing Farm Products,* 7th ed. (Ames, Iowa: Iowa State University Press, 1982).

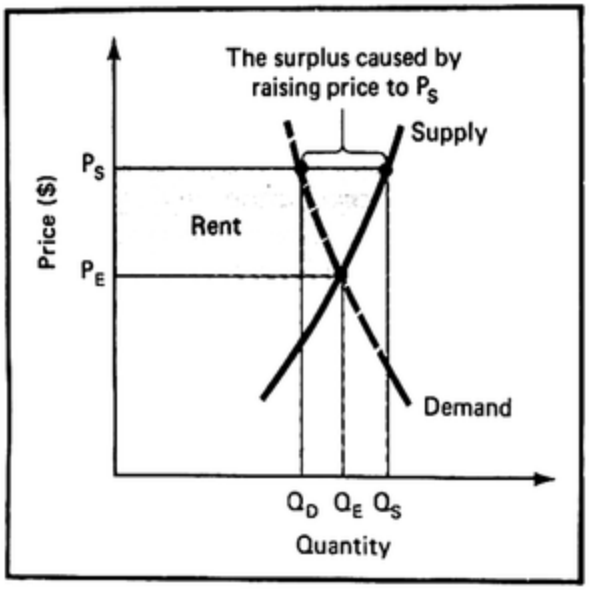


Figure 7 The main effects of raising farm prices

Long-run equilibrium is reached at P_e and Q_e. The supported price P_s results in consumers buying Q_D while farmers produce Q_s. The difference ($Q_s - Q_D$) is a physical surplus. Although it is "only" about 30 percent of the crop (because both demand and supply are inelastic), the yearly surplus can become enormous as the years pass.

program, efforts were made in the 1950s to reduce them while still keeping farm prices up. To reduce output, farmers were paid to take some of their land out of cultivation and put it in a "land bank." This was costly, the more so because farmers naturally set aside their worst land. Moreover, they tilled the remaining land more intensively because their other inputs (equipment, labor) were little changed. Accordingly, total production was little reduced. The land bank, a clear case of economic waste and a costly failure, was abandoned in the 1960s.

Surpluses were also channeled into school milk programs and giveaways to needy foreign countries. But such "Food for Peace" types of foreign aid depressed farm prices and farm incomes in those countries receiving the foods, thus harming farmers and undermining the prospects for agricultural progress in those countries.

Certain farm prices were also raised indirectly. For example, the land permitted for growing tobacco has been rigidly limited since 1938. That has raised tobacco prices and enriched landowners by raising the price of the land far above its economic value. The same is true of peanut acreage. About 40 other crops (from dates and grapes to California oranges) are under "marketing orders." When their prices threaten to fall below target prices in bountiful crop years, the Agriculture Department buys the surplus and destroys it.

Such waste could be avoided and needy farmers could still be helped—but only by avoiding any effort to change farm *prices*. The efficient treatment would simply identify the needy farmers and raise their incomes by direct payments. The payments would be based on an appraisal of the farmers' true degree of need, just as welfare programs are mainly confined to those who are genuinely in need. The payments would, therefore, flow mainly to small-scale farmers, as in Appalachia and parts of New England. Little or nothing would go to the big commercial farmers on the plains, who are not needy.

Being relatively focused, the payments would be much less costly to the public purse than programs based indiscriminately on prices and outputs. They would avoid enriching already-rich farmers and would not raise farm prices. They would also avoid needless economic rents and windfall gains from rising land prices.

The only drawback is that direct payments might encourage people to stay on farms that are too small and/or barren to justify farming. Such marginal farms should, by strictly economic criteria, be abandoned or merged into larger farms. Direct income payments might tend, instead, to encourage a permanent class of small operators on inefficient farms.

Yet, that mistake could easily be avoided. The direct payments could be

kept low enough to encourage inefficient farmers to migrate to towns. Or specific relocation grants could be provided to encourage marginal farmers to shift toward more productive jobs and locations.

By contrast, actual farm programs based on farm prices have had sharply inefficient and unfair results. Poor marginal farmers have received little benefit, and the prices of farms have been driven beyond their reach. Since the 1930s, that has forced many small farmers off the land.

The whole topic illustrates how economic analysis can clarify complex situations, even with simple tools. It also shows how policies can continue in error long after their economic faults are exposed.

The economics of energy

The modern economy uses vast amounts of energy in farms, factories, homes, and transportation. Moving from a primary reliance on wood in the early 19th century, to coal and finally oil in this century, economic expansion has absorbed rapidly growing volumes of energy. The 1970s saw an abrupt end to a long period of energy abundance in which coal, oil, and gas were discovered in ever growing amounts. Industrial economies are having to make severe adjustments to cope with the rising energy scarcity. Energy economics has become an urgent topic.

Basic trends

We begin by adopting a large perspective, for the current issues are best seen as small parts of fundamental shifts. The modern dependence on fossil fuels—especially coal, oil, and natural gas—is a brief episode in human history. Since they took 300 million years to form, these resources will not be replaced in any meaningful degree. Even the vast reserves of coal will largely be gone in a few centures.

As for oil, the United States will probably pass its peak rate of production in the 1980s; world production will probably peak by 1995, as Figure 8 shows. Indeed, the Middle East oil kingdoms expect to exhaust most of their reserves in about 30 years.If you are 20 years old now, the decline may begin by your 40th birthday; when you are 65, some 80 percent of all recoverable world oil reserves will have been consumed. Economic growth will have to be sustained somehow on a shrinking flow of oil.

The economic use of fuels begins with the easiest and cheapest sources and then moves to less accessible costlier reserves. That exactly parallels the use of the best farmland first, the best hydroelectric sites, and every other natural resource. As the best coal is used up, shafts must be sunk to deeper veins. Similarly, oil and gas must be sought in ever more inaccessible places. *In short, the use of fuels proceeds up a rising scale of costs.*

The present array of sources ranges from shallow Mideast oil wells to expensive, capital-intensive solar equipment. As the actual prices of fuels rise, the margin of production from these sources shifts to the right. If oil sells for $20 per barrel, shale is not worth processing to produce oil. But if the price of oil rises to $40, then massive oil shale mines and processing plants may become economic. *The whole issue turns on the future opportunity costs of alternative fuel sources.*

Note that the rising scarcity of fuel is a matter of degree. Fuel reserves are not a single pot that contains a definite amount. The *physical* presence of fuels is ultimately a given: Whatever is there is there. But, *economically*, there is a wide spectrum of choices, moving from cheap fuels to very costly ones. Rather than "run out" all at

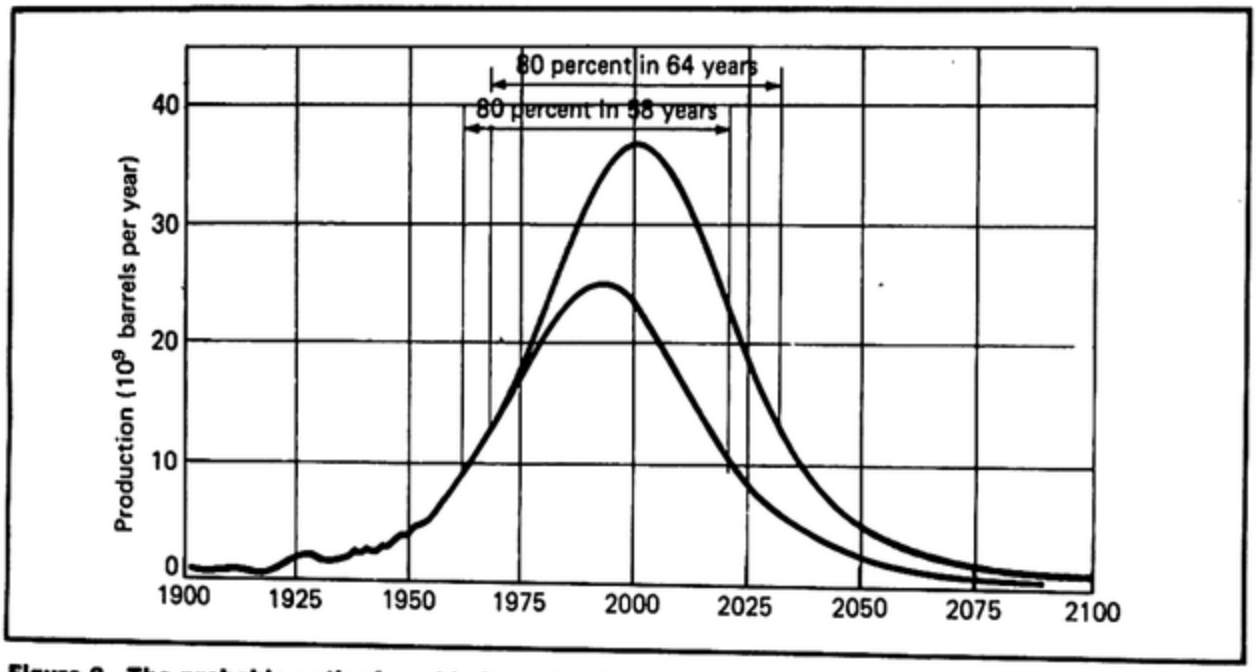


Figure 8 The probable path of world oil production

The cycle of world oil production is plotted on the basis of two alternative estimates of the amount of oil that will ultimately be recovered. Even by the larger estimate, 90 percent of the recoverable oil will be used by about 2030. Even a marked slowing in consumption will only postpone the scarcities by a few years.

Source: Adapted from M. King Hubbert, "The Energy Resources of the Earth," *Scientific American*, September 1971.

once, the world will move to increasing scarcity, which will take the form of rising energy prices. We explained that process in the first main section of this chapter. The world has been "running out" of the best energy sources for a long time. This process will continue, probably pushing energy prices even higher—unless technology creates new, cheaper sources.

Indeed, energy prices do not rise only when the supply dwindles. Investors will realize the coming scarcity and try to buy the reserves now, which in itself will send up prices. ***In short, the market anticipates the physical shortages, building the expected future price increases into the current price of fuel.*** That ensures that fuel prices will rise as soon as future scarcity is adequately perceived. On both sides of the market, economic processes make price trends reflect and anticipate true long-run scarcities.

In retrospect, then, the fuel gluts of the 1950s and 1960s —when oil and gas prices stayed at relatively low levels—were a marked oddity. They reflected the brief flooding of world markets by vast new Middle Eastern oil reserves. There were also bright hopes for cheap nuclear power. But those were just fluctuations around the trend. Though they did lull many shortsighted analysts into believing that fuels would always be abundant and therefore cheap, the 1970s brought an abrupt return to the reality of advancing scarcity.

Since most fuels can be substituted for one another in at least some uses, the rise in oil prices naturally induced a parallel rise in other fuels, including coal, gas, nuclear fuel, and even firewood. All of these fuel price rises caused windfall gains for the owners of the fuels, as rising prices were capitalized into the value of the fuel reserves. There was also a worldwide cartel in nuclear fuels from 1969 to 1974, which attempted to fix the price of uranium above competitive levels. But the net effect of that cartel is debatable, for by

1973, rising oil prices were already causing uranium prices to rise strongly.

Indeed, even OPEC's price fixing may not have raised the price of oil much higher than it would have been pushed by market forces. The rising scarcity of fuels would have caused oil prices to rise naturally in any event, as owners and investors bought reserves in anticipation of future price rises.

OPEC seemed to force the price up sharply during the Arab oil embargo in 1973–1974, and the Iran-Iraq war seemed to cause another sharp rise in 1979–1980. But they may have had little real effect. One concludes that the underlying pattern was as shown in Figure 9. If it is valid, then OPEC merely moved oil's price up in 1973–1974 a little earlier than it would have risen from the other causes. The rest of the rise has not been caused by OPEC. Other experts blame all of the price rise on OPEC and suggest that conflicts within OPEC will cause oil prices to fall back to $20 per barrel or below.

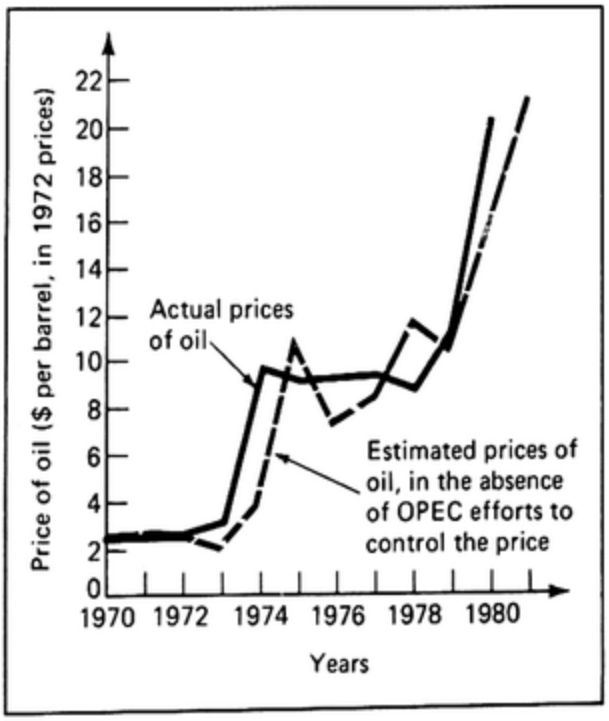


Figure 9 Has OPEC really raised the long-run price of oil?

MacAvoy's projection suggests that OPEC did not affect the long-run upward trend in oil prices, but only moved it ahead in 1973–1974. However, other specialists suggest that OPEC intensified the rise, even if it was not the sole cause. The price of oil is stated here in constant 1972 dollar values, in order to filter out the effects of general inflation.

Based on Paul W. MacAvoy, *Crude Oil Prices: As Determined by OPEC and Market Fundamentals* (Cambridge, Mass.: Ballinger Publishing Company, 1982). Reprinted with permission.

Seen in perspective, rising oil scarcities probably increased oil prices strongly after 1970, perhaps as much as MacAvoy suggests. The debate is not about the size of OPEC's vast oil revenues, but about how those money flows are defined: either as rents (if oil prices would have risen anyway) or as monopoly profits (if OPEC has raised the prices artificially). The debate is even more important for what it leads us to expect about future oil prices: If the long-term trend is up (reflecting a rising global scarcity), then further rapid rises are likely.

But if OPEC's price fixing has been the real force raising oil prices, then OPEC might not be able to raise them further or even to maintain present price levels. If the sharp differences among OPEC countries causes them to disagree about the joint-maximizing price (recall Chapter 11's analysis of collusion), their price-fixing efforts might collapse altogether, or at least stabilize at a much lower price level. Most likely, the true answer involves a mixture of these two points of view.

The search for oil Whatever its causes, the steep rise of oil prices has stimulated the search for more oil deposits. Much of the effort has been focused on the ocean floors, via hundreds of large oil rigs, but discovery has also moved into remote regions. The search for gas has also been spurred by the removal of some controls on the price of U.S. gas. These added efforts have brought some results. But the marginal returns to exploration continue to decline. That is inevitable, for the cheapest, most accessible sources were ex-

ploited decades ago. Moreover, the more remote oil and gas reserves are also costlier to transport to the market. For example, oil and gas pipelines from Alaska are large-investment projects, which have had to overcome severe problems of climate and terrain.

All of this perfectly fits the logic of rising long-run scarcity, but ultimately, of course, no discovery, however fast, can increase the amount of fuel in the ground by so much as a gallon. It only confirms the reserves' existence sooner and makes it possible to convert them to human uses more rapidly. Indeed, a rapid discovery and use rate may *violate* the long-run conservation of fuel, which would call for it to remain stored in its natural forms to be used later. That choice, in turn, depends on the long-run prospects for alternative fuel sources, such as solar energy, nuclear power, fusion processes using plain water, and ocean tides. Those future trends are mostly unknowable now. Only if other sources are going to be cheap and abundant would the rapid discovery and use of oil be efficient now. If, instead, the prospects for new technology are dim, then oil prices should rise more rapidly now, causing the use of new oil to be postponed rather than hastened.

Future world resources

Fuels are not the only resources that appear to pose severe long-run economic hazards. Other threatened resources include many ores, chemicals, topsoil, animals, and water sources. All of these are threatened by population growth and economic growth.

The population on the earth has risen to nearly 5 billion people, from only 1 billion in 1830. Despite a recent decline in the growth rate, especially in the industrialized economies, the world's population is certain to continue to rise substantially; the only question is how far. The answer depends on the *replacement-level birth rate.* It is about 2.2 children per woman, which is above current levels in industrial countries but well below current levels in the more populous developing countries. Even if a replacement-level rate is achieved, a momentum factor will operate to keep the population rising for several decades more.* Accordingly, the following outcomes can be predicted:

If the world attained replacement fertility in:	World population would then stand at:	And world population could be expected to stabilize eventually at:
2000–2005	5.9 billion	8.4 billion
2020–2025	8.4 billion	11.2 billion
2040–2045	12.0 billion	15.1 billion

World population will almost certainly double because replacement fertility cannot be reached by the years 2000–2005. Indeed, few experts are confident that it will ever be attained. Even if it is reached as soon as 2020–2025, population will still grow to 2.5 times the present 4.5 billion.

To feed, shelter, and clothe the added population will strain the world's resources of space, housing capacity, and food production. Moreover, the added population will add to the demand for products using fuels, ores, chemicals, and other exhaustible resources.

To accentuate this pressure, future populations will want higher real incomes than people today have. Even if real in-

*The numbers of young people coming into the age of fertility in developing countries are so much larger than the numbers of people at older ages that, even if couples have only enough children to replace themselves, total births will continue to outstrip total deaths until the disproportion in the number of people in the childbearing ages disappears. This could take a century or more in countries where population growth has been very high. The size of the population will be far larger than when fertility dropped to the replacement level.

comes per capita increase by only 1 percent per year and population stabilizes at the "low" level of 8 billion, total production will need to rise by more than 100 percent by 2020 and by an another 100 percent by the year 2050.

This growth will have to occur despite shrinking natural resources. As some resources become more scarce, their relative prices will rise, perhaps drastically. Nations lacking those resources will face rising economic pressure. International tensions may become severe, with chronic conflicts over the control of valuable oil, ores, fish, chemical, and agricultural resources. Already there have been conflicts over ocean fishing rights, fertile land, and oil fields. Seizure of land and resources has been a common cause of warfare throughout human history, and the pressures for such wars will intensify. It is a dismal prospect: increasing international stress as resources dwindle and population rises.

Yet, economists have largely rejected this doomsday view, for reasons that reflect the very nature of economics. The doomsayers rely mainly on engineering models, with fixed relationships among the causes and effects. They assume that certain key resources, including oil and ores, would continue to be used at past rates. Projected forward by 40 or 50 years, those rates truly could not be sustained, and some sort of collapse seemed inevitable.

But the models left out the key role of *prices* and *elasticities*. Rising scarcities will be reflected in rising prices. Both demand and supply may respond to the changing relative prices, so that the rates of physical usage do not charge ahead at a constant rate. On the contrary, as the oil price rises of the 1970s showed, the demand for fuels has substantial elasticity, especially in the long run. Such elasticities may be reasonably high for most or all of the crucial resources.

Therefore, economies are likely to adjust their resource use, softening the impact of resource scarcities. Rather than rush blindly into radical physical shortages, the economies are likely to absorb the changes gradually and minimize their impact. The doomsday predictions were therefore said by most economists to be too pessimistic: The disasters are conceivable but not likely. There will be soft landings, they said, not crashes.*

Summary

The main points have been:

1. Each resource has an optimal rate of use, which maximizes the net present value of all future uses. This rate reflects a discounting of benefits and costs. Market choices may reach the optimal rate.
2. Common-property resources will be used too rapidly by free-market activities. Other possible causes of non-optimal use include wrong discount rates and external effects.
3. Farm policies usually cause inefficiency when they attempt to raise prices rather than to stabilize farm incomes.
4. The rise in energy prices in 1970 reflected both long-term shifts and, possibly, OPEC's raising of oil prices.

*Julian Simon goes further, arguing that added population will increase the capacity to produce and to solve global problems. He also predicts that technology will adjust to resource scarcities and maintain the long-term growth of per capita incomes. The view rests on optimistic expectations of new technology and human capacities, in the tradition of technological optimists who "refute" the Malthusian predictions. See Julian L. Simon, *The Ultimate Resource* (Princeton, N.J.: Princeton University Press, 1981).

5. Future resource scarcities may cause world-wide crises within several decades. But price changes and elasticities can make for smooth rather than disruptive adjustments.

Key concepts

Conservation
Discounted net present value
Optimum rate of use
Intergeneration choices

Questions for review

1. a. What is the rule for determining the optimum rate of use of an exhaustible resource?
   b. List the five main elements which determine the optimum rate of use of a resource. Discuss how each of them influences the optimum rate of use.
2. Explain how each of the following may interfere with the most efficient use of resources:
   a. common-property resources,
   b. political influences,
   c. distribution of wealth and income.
3. Farm programs have moved from income stabilization to price raising. Discuss two economic effects of this change in goals.
4. Using economic analysis, explain how an increase in oil prices could lead to an increase in the cost of firewood.

22

An Introduction to Macroeconomic Analysis

As you read and study this chapter, you will learn:

- that in the long run our economy grows at 3 or 4 percent a year, about half of which is a gain in per capita output
- that the growth is not steady, but punctuated by alternating periods of prosperity and depression
- how this cycle of boom and bust is reflected in the unemployment rate
- that in the past, long periods of falling prices were nearly as common as periods of rising prices
- how the 1970s were unique in having both high unemployment and inflation at the same time

At its southern tip, the eastern seaboard doesn't really end, it just trails off in a string of islands that sink into the Gulf of Mexico like the words of an unfinished sentence. These are the Florida Keys. A few miles from the Keys is a parallel string of islands, the famous offshore reef. These boast some of the most spectacularly beautiful scenery our country has to offer, but it is entirely underwater. If you go there, you can view it from a glass-bottomed boat. But to appreciate fully the richness of form and color of this remarkable formation, you really have to dive deeply enough to look up at it from the bottom of the ocean.

Apart from its sheer beauty, the reef is fascinating because it is really an enormous colony of microorganisms known as coral. The offshore reef is a social formation, not a geological one. Its fascination lies not in the individual lives of its members, which must be stupefyingly dull, but in the monumental outlines of its collective life process.

The collective life processes of humanity are no less fascinat-

ing, nor are they less beautiful in their form and color. So take a deep breath and prepare to dive into ***macroeconomics,*** the study of the collective behavior of the microorganisms that make up the American economy. Even if your interest in economics did not extend beyond your own well-being, it would pay you to take an interest in macroeconomics. Unless you are a hermit, producing for your own subsistence with tools fashioned by your own hand, your economic life is hopelessly entangled with those of others. This is true of everyone. We are all enmeshed in an *economic system.* This means that our individual lives are largely governed by fluctuations in the national economy.

This chapter is a brief introduction to trends in the national economy. Its main purpose is to put the chapters that follow in perspective, to enable you to see what they explain. The chapter is divided into two main sections. The first covers the outlines of more than a century of economic development, from shortly after the Civil War to the present. Although the figures presented are only a silhouette, they show clearly a pattern of long-term growth punctuated by short-term instability. The second section looks at the decade of the 1970s in detail. It shows that the economy-wide trends and instability are shared by many of its sectors, so that what is true for the whole is also true for many of its parts.

Long-term trends, 1870–1980

Gross national product

This chapter is organized around a series of diagrams, the first of which (Figure 1) shows national output over the last century. The series on the graph is ***gross national product,*** or *GNP. GNP measures the national output of all goods and services that are bought and sold on the market*—consumer goods; business purchases of plant, equipment, and inventory; government purchases; and exports, less those goods that are imported. The chief impression Figure 1 conveys is that output has gone up and up during most of this period. Since output is measured in dollars of constant purchasing power, this is genuine growth, not simply price increase. A trend representing a steady growth rate of 3¼ percent a year is included as a benchmark for comparison. This gives the right order of magnitude of the growth trend in GNP, although you can see that over some fairly long periods growth has sometimes been faster and sometimes slower than that. Since the population growth rate over the same period was only 1¾ percent a year, a good bit of the overall GNP growth translated into higher per capita output. In fact, output per capita increased about sixfold in these 110 years. This has meant an enormous increase in living standards and national wealth.

After the pronounced upward trend in output, the second thing you probably notice is the year-to-year fluctuations relative to that trend. This is usually called the ***business cycle,*** the rhythm of good times and bad times. Some particular historical periods are flagged on the graph, so that you can relate the business cycle to events in your and your family's memory, and to things you have read about in history and literature. The most prominent of these periods is the ***Great Depression,*** extending from the stock market crash in the fall of 1929 until the beginning of World War II, more than a decade later. This was a worldwide economic disaster, affecting nearly every part of the world economy. It was far deeper and more prolonged in the United States than in the rest of the world, and caused more misery, hopelessness, and moral outrage at the social order than any other event in our modern history. Few people who were affected by the Depres-

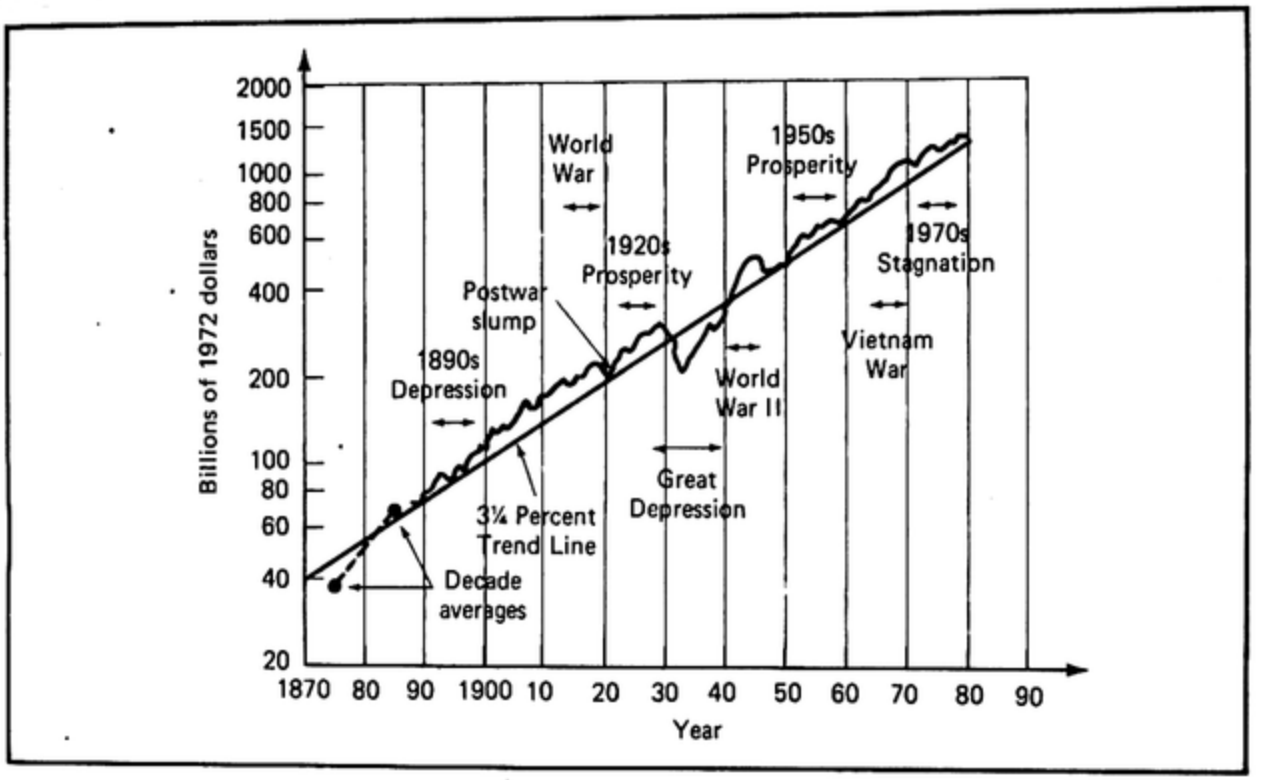


Figure 1 GNP 1870–1980 (in constant 1972 dollars)

This diagram shows the history of gross national product or GNP, which is the national output of goods and services bought and sold on the market, from the 1870s through 1980. The vertical scale used on this and many later diagrams is a logarithmic or ratio scale, on which equal percentage changes appear as equal distances. This means that steady percentage growth per year can be graphed as a straight line. As a basis for comparison, the diagram includes a 3¼ percent trend line. As you can see, this trend line approximates the upward drift in GNP itself. Since GNP is measured in constant dollars, this is real growth, not simply price increase. However, as the diagram also indicates, GNP has fluctuated around this trend in a pattern that is often called the business cycle.

Long-term comparisons such as this are subject to substantial measurement problems. Data for the 1870s and 1880s are so rough that only the 1870–1879 and 1880–1889 averages are shown. Data from 1890 to 1929 are not very good either. They do give the right order of magnitude, however. GNP has doubled about every 20 years, not every 2 years or every 200 years.

Source: U.S. Department of Commerce, *Long-Term Economic Growth; Economic Report of the President.*

sion escaped it without some emotional wounds and bitter memories. It enriched our vocabulary of misery with evocative words like soup kitchen, bread line, dust bowl, Okie and Arkie, Hooverville, CCC camp, and WPA project. If you don't know what these mean, ask your grandparents or read the novels of John Dos Passos and John Steinbeck—and hope that it does not happen again in your lifetime.

Unemployment

The hallmark of depression is unemployment, and fluctuations in unemployment tell the story of the business cycle in terms more immediately human than fluctuations in national output. That story can be read in Figure 2. ***The unemployment rate is the percentage of the labor force that cannot find work.*** You may be staggered to see how high it was during the 1930s. It peaked at 25 percent in 1933 and didn't drop much below 15 percent until the end of the decade. These figures understate the numbers of workers directly affected. Two people out of work for half a year each only count as one unemployed. People whose work week was cut to 20 or 30 hours aren't counted at all. Nor are those who didn't bother to look for work because they knew none was to be found.

When compared with the 1930s, the unemployment rates of recent decades

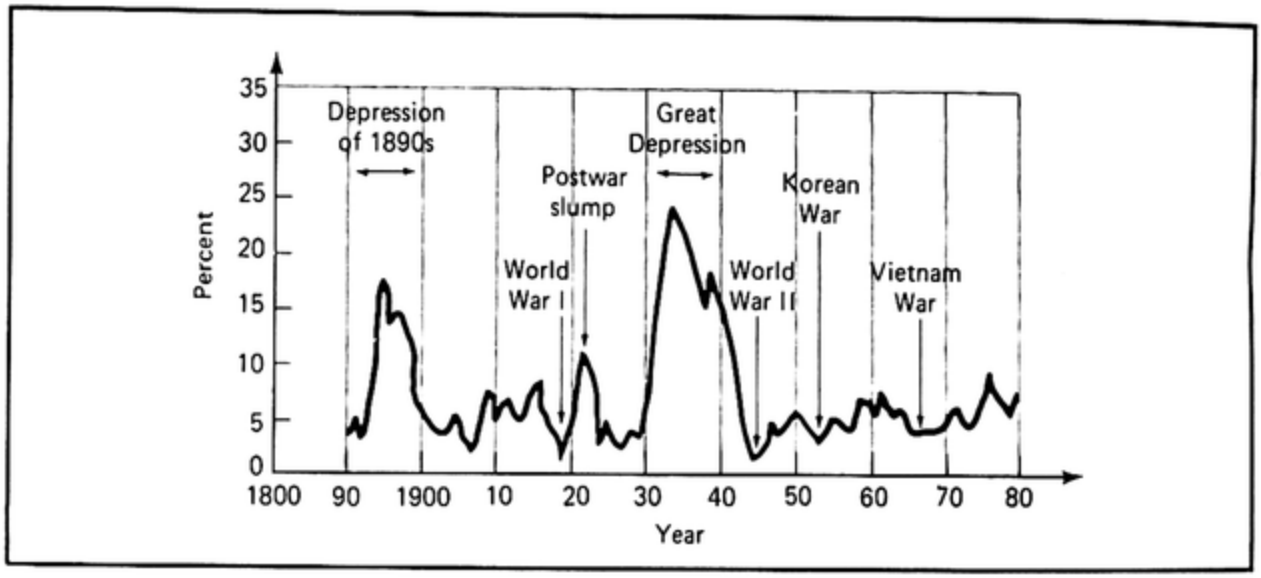


Figure 2 The U. S. unemployment rate 1890–1980

The business cycle fluctuations in GNP are mirrored in the unemployment rate, which is the percentage of the labor force unable to find work. As you can see, unemployment during the 1890s and especially during the Great Depression of the 1930s dwarfed anything we have seen since. But the average unemployment rate in 1975 was 8½ percent, its highest level until then since the 1930s.

Source: U.S. Department of Commerce, *Long-Term Economic Growth; Economic Report of the President.* Figures before 1929 are estimates prepared by Stanley Lebergott and are published with his permission.

seem innocuous. But consider some figures: In 1975, the unemployment rate averaged 8½ percent, its highest point until then since the Great Depression. In *each month* of that year, there were nearly half a million *new* applicants for unemployment benefits. If that many people had to get in a single line to get their benefits, it would reach from New York to Washington. Moreover, some groups suffered much more than others: The unemployment rate among white teenagers was about 18 percent, among black adults about 12 percent, and among black teenagers nearly 40 percent.

Consumer prices

The consumer price index *measures changes in the average prices paid by a typical consumer.* If the index is 100 in 1967 and 110 in 1969, this means that on average consumer prices rose 10 percent over the two intervening years. Figure 3 shows the history of consumer prices from 1860 through 1980. You may find it hard to believe, since it shows several periods of falling prices. You have probably never known anything but rising prices.

Both the Civil War and World War I were followed by periods of falling prices. Both the depression of the 1890s and the Great Depression of the 1930s also produced sharp declines in the price level. But prices rose after World War II, except for brief lulls in 1949 and after the Korean War. In the main, sharply rising prices have resulted from wartime and immediate postwar shortages. Most of the big swings, whether up or down, have reflected the peculiar circumstances of war and its after effects.

The record of peacetime fluctuations in prices is mixed. During the relatively peaceful half-century between the Civil War and World War I, the trend was mostly downward. Prices were stable during the prosperous 1920s, which intervened between the post–World War I and

Depression price drops. Since World War II, prices have generally risen.

The big exception to the pattern of price fluctuations that characterizes most of Figure 3 is the 1970s. The 1970s were peaceful but hardly prosperous. Past experience predicted that prices in the 1970s should have remained steady or even dropped. Yet you surely don't need Figure 3 to tell you that the high unemployment of these years was accompanied by an exceptionally rapid rise in prices, which more than doubled over the decade. Because of this, recent economic history has been fascinating, if painful, to live through.

Common stock prices

The stock market is just one among many financial markets, and the amount of capital raised on this market is not very large relative to that raised on the others. Nonetheless, the stock market is generally thought to be especially interesting because of what it has to tell us about the state of "investor confidence." Common stocks entitle their owners to a share of future profits of the companies that issue them. But nothing guarantees that the profits will ever materialize. Bonds are promises to pay definite sums at definite dates unless the issuers are bankrupt. But stock dividends are an "iffy" proposition. When financial investors change their expectations of future profits, they change their valuations of titles to a share of these profits. The market reflects these changes, which is why it is such an interesting barometer.

If you compare Figure 4's history of stock prices with the output fluctuations shown in Figure 1, you will see that they have a lot in common. Stock prices rise in good times and fall in bad. This is neither surprising nor very interesting. What are interesting, however, are the differences between the two diagrams. The one that stands out most sharply is the spectacular stock market climb of the 1920s, followed by the plunge of the 1930s. Stock prices increased nearly fourfold from 1921 to 1929, although GNP rose by only 60 percent. People of quite ordinary means made considerable gains by borrowing to buy stock

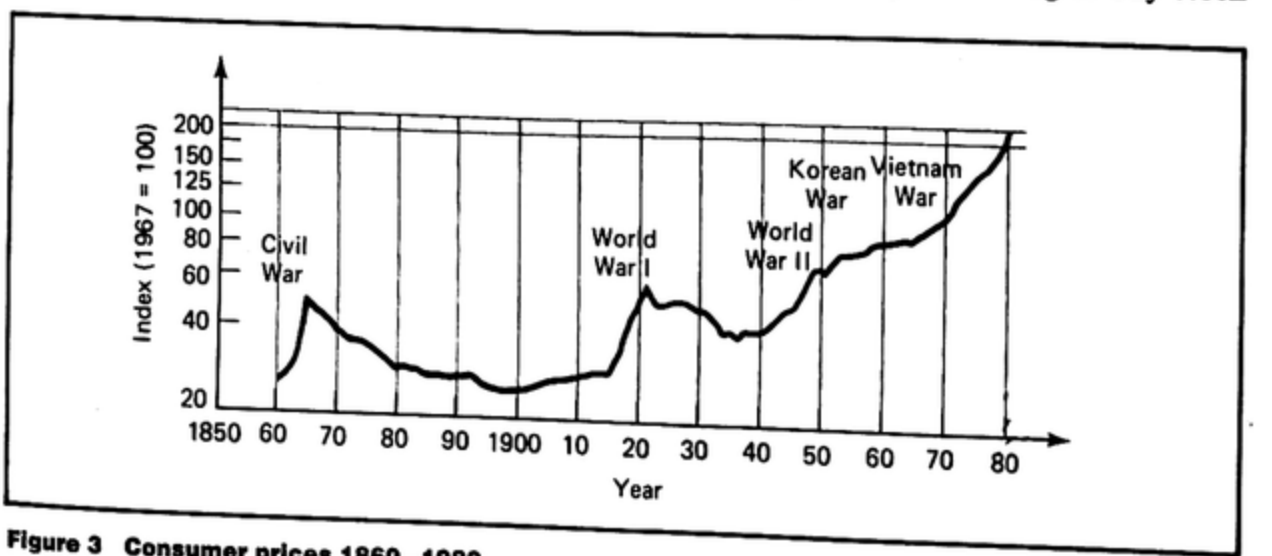


Figure 3 Consumer prices 1860–1980

The consumer price index measures changes in the average prices paid by a typical consumer. If the index is 100 in 1967 and 110 in 1969, this means that, on average, prices rose 10 percent over the two intervening years. As you can see, prices have gone both up and down since 1860, but have moved predominantly upward since World War II.

Source: U.S. Department of Commerce, *Long-Term Economic Growth; Economic Report of the President.*

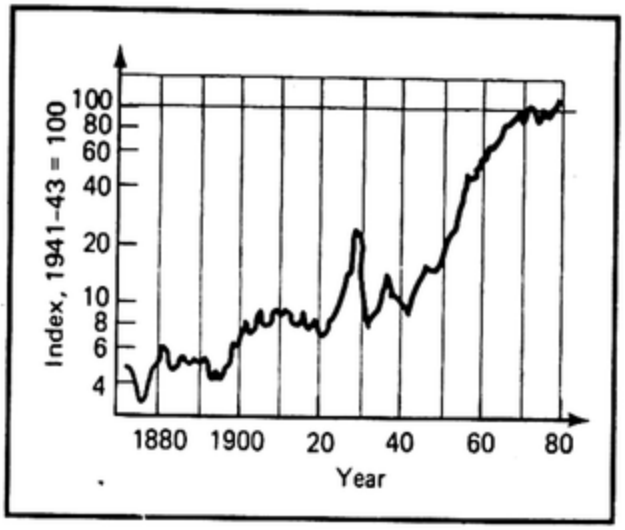


Figure 4 Common stock prices 1870–1980

Most of the time, the fluctuations in common stock prices follow fluctuations in GNP. The exceptions are more interesting than the rule, however. Notice particularly the enormous rise in stock prices in the 1920s and the collapse in the early 1930s. Both rise and fall were quite out of proportion to the fluctuations in GNP. Notice also the stagnation of stock prices during most of the 1970s, despite the expansion in GNP.

Source: U.S. Department of Commerce, *Long-Term Economic Growth; Economic Report of the President.* Figures prior to 1908 were prepared by the Cowles Foundation; thereafter by Standard and Poor's Corporation.

in an orgy of speculative fever. Nothing in the general prosperity of the 1920s justified the lofty peak to which the market rose. What started as an expression of confidence in the future profitability of American capitalism turned into an expression of confidence in rising stock prices. This is the essence of a speculative bubble. It is held up by a belief that it will continue to stay up—nothing more. When the stock market broke in October 1929, it began a plunge that carried it back to 1922 levels, wiping out the fortunes it had created in the 1920s. Although the basis for the upswing was insubstantial, the bankruptcies caused by the crash were real enough. They surely contributed to the cutbacks in spending responsible for the decline in real output.

Another interesting episode in the history of the stock market is the leveling off of stock prices during most of the 1970s after two decades of nearly steady growth. Despite two recessions, national output grew at a rate of 2.8 percent a year from 1968 to 1978, not all that far below the long-term trend. Why, then, were stock prices so stagnant? No one knows for sure, of course, but several factors may have been responsible for the general pessimism. The first is the unique combination of high unemployment and rapid inflation that gave rise to the word *stagflation.* If you carefully compare Figures 2 and 3, you will see that periods of above-normal unemployment have generally produced steady or falling prices. This led people to believe that a tradeoff existed between inflation and unemployment. Somehow the economy and its policymakers had to choose the right combination, but it seemed possible to have more of a good thing, say high employment, if we were willing to put up with more of a bad thing, inflation. The 1970s seemed to indicate that some profound change had occurred, that we were destined to have more of both social ills. On top of this, we had the energy crisis, the Iranian crisis, the Afghanistan crisis, one after another until crises seemed to be the normal state of affairs. In view of all this bad news, it is less surprising that people came to wonder if uncertain future income from stock ownership was worth a high present price.

The federal government

A major structural change in the economy has taken place since 1929: an enormous growth in the federal government relative to the rest of the economy. Figure 5 illustrates one of the dimensions of this growth, the expansion of that part of national output purchased by the federal government. Despite the lack of a definite uptrend between the Korean War and the end of the 1970s, the overall growth rate of federal purchases from 1929 to 1980 was

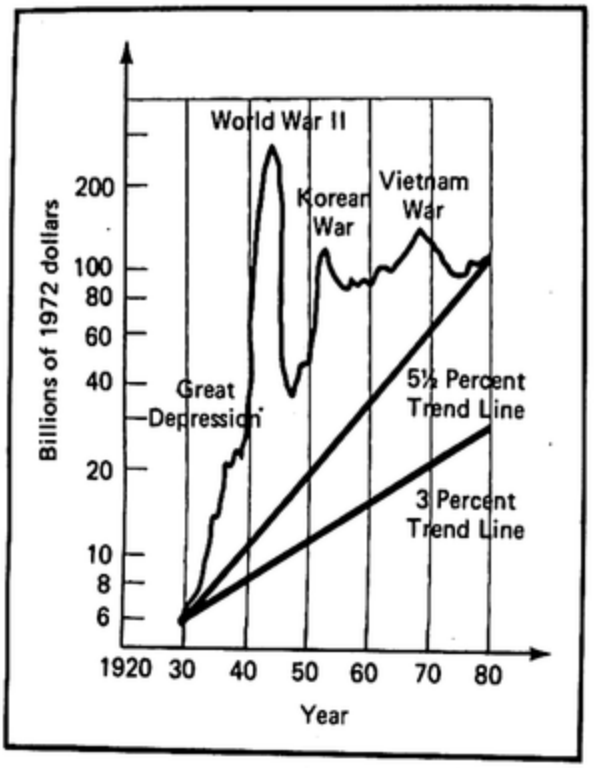


Figure 5 Federal government purchases of goods and services, 1929–1980 (in constant 1972 dollars)

From 1929 to 1980, federal government purchases of goods and services grew at about 5½ percent a year in constant dollars. Over most of the period, it was way above that trend. The 3 percent trend line represents the average growth in national output over the same historical period. (This trend from 1930 to 1980 was slightly below the long-term trend in GNP.) If federal purchases had merely kept pace with national output, they would only have been a little over $30 billion in 1980.

Source: *Economic Report of the President.*

5½ percent a year in constant dollars. Over the same 51 years, national output grew at an average rate of 3 percent. As a result, government spending was three times as big in 1980 as it would have been had it remained a constant share of national product.

It is generally believed that this great relative expansion of the federal government has been mainly associated with wars. As you can see from Figure 5, the start of each of the three wars since 1929 was accompanied by a rapid rise in federal expenditures, especially, of course, during World War II. It is equally true, however, that the end of each of these saw a rapid drop in military expenditures. The long-term expansion in the relative importance of the federal government took place in two stages. The first was during the Great Depression and was concentrated in expenditures for civilian purposes. By 1940, federal nonmilitary purchases had reached a peak relative to GNP that has not been exceeded since. The second was during the late 1940s and, especially, the 1950s, when peacetime military expenditures ballooned during the early years of the Cold War. To a degree, this was reversed during the "détente" of the 1970s, but the peacetime military establishment of 1980 absorbed 5 percent of GNP, more by a factor of 3 than the entire federal government absorbed in 1929.

The size of the surplus or deficit in the federal budget is closely related to the expansion in the relative importance of the federal government. When government expenditures exceed tax receipts, the budget shows a deficit. When receipts exceed expenditures, it shows a surplus. Figure 6 depicts the federal surplus or deficit as a percent of GNP. Obviously, it is dominated by

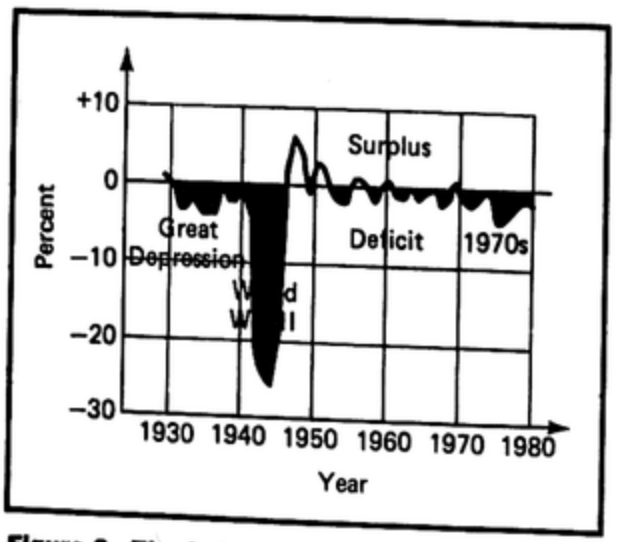


Figure 6 The federal surplus or deficit as a percent of GNP 1929–1980

The historical pattern of the federal deficit is dominated by the World War II period. Outside this period, chronic deficits were incurred in the 1930s and the 1970s.

Source: *Economic Report of the President.*

Alternative Views on the Stability of Capitalist Economies

Diagrams like Figures 1–4 are not self-interpreting. Marxist economists who look at these diagrams will see a pattern of repeated and disastrous instability. To them, the Great Depression was just the worst of many similar episodes. Although they see the capitalist system as a great engine of growth, they are not particularly impressed with the tendency for unemployment to remain within fairly narrow bounds most of the time. They view capitalism as an intrinsically self-destructive social organization that will one day disintegrate from an overdose of its own contradictory tendencies.

A second group views capitalism as a system with flaws that can be remedied by carefully administered applications of their own brand of economics. They recognize the same pattern of instability seen by the Marxists, but think they know how to do something about it. These are the Keynesians.

A third group is much impressed by the uptrend in national product per capita and by the ability of the American economy to keep its instability within bounds. They trace the episodes of instability, even the Great Depression, to faults in the structure and regulation of the country's banking and monetary system. They believe that if the nation's money supply could only grow at a steady, moderate rate per year, most of the instability would disappear. These are the *monetarists*. The principal spokesman of the monetarists is Professor Milton Friedman (1912-), who was for years the dominant voice of the economics faculty at the University of Chicago.

Three Economists Who Have Shaped Our Views About Economic Growth and Instability



KARL MARX



JOHN MAYNARD KEYNES



MILTON FRIEDMAN

World War II, when the federal deficit was enormous because the government chose to finance the war by borrowing rather than by taxing. As for the rest of the period, careful reading will show that the budget moves toward surplus in prosperous times and toward deficit in depressed times. You can see a pattern of chronic deficits during the Great Depression. You can also see another period of chronic deficits during the troubled 1970s. This was an important—and very troubling—part of the

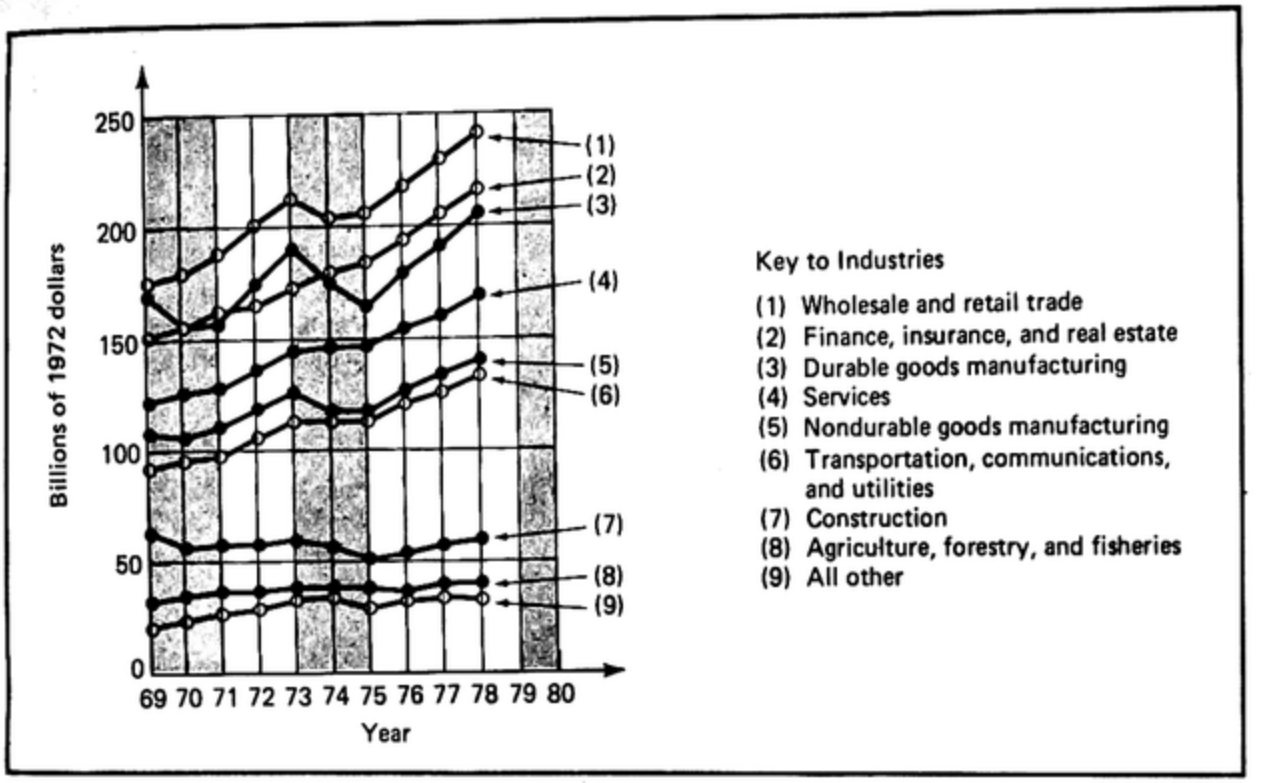


Figure 7 Output of U.S. private industries, in constant dollars 1969–1978

Output fluctuations in most industries mirror those in total GNP.

Source: *Economic Report of the President.*

overall pattern of developments during this period.

There are a number of conflicting views on the significance of the expanded role of government. Some of the main ones are outlined in the box in this chapter. After you have read the next 10 chapters, you should be in a better position to make up your own mind on the matter. But no matter what a person may think about these issues, he or she must recognize that our economy is influenced by the federal government to a degree that simply was not true 50 years ago. The analysis of growth and instability could at one time be concerned almost entirely with the behavior of private individuals and institutions. Now it must also study their interactions with a large and powerful government.

Cyclical fluctuations in the 1970s

One very important dimension of cyclical fluctuations cannot be seen in the diagrams you have looked at so far. This is the extent to which the rhythm of the business cycle sets the tempo for nearly every sector of the economy. This section of the chapter remedies this shortcoming by looking at how production, unemployment, and consumer and stock price variations affect various industries, products, and groups of people.

Production

The gross national product, whose long-term fluctuations and growth you looked at in Figure 1, can be broken down into the products of the various industries that collectively make up our economy. Figure 7 does this, presenting 10 years of business fluctuations as they were recorded by American private industry. The business cycle may be conveniently divided into periods of *expansion* in output, ending at a cyclical *peak*, followed by periods of *contraction* in output, ending at a cyclical *trough*. Peaks and troughs are usually located to correspond to the low and high points in the unemployment rate. You will

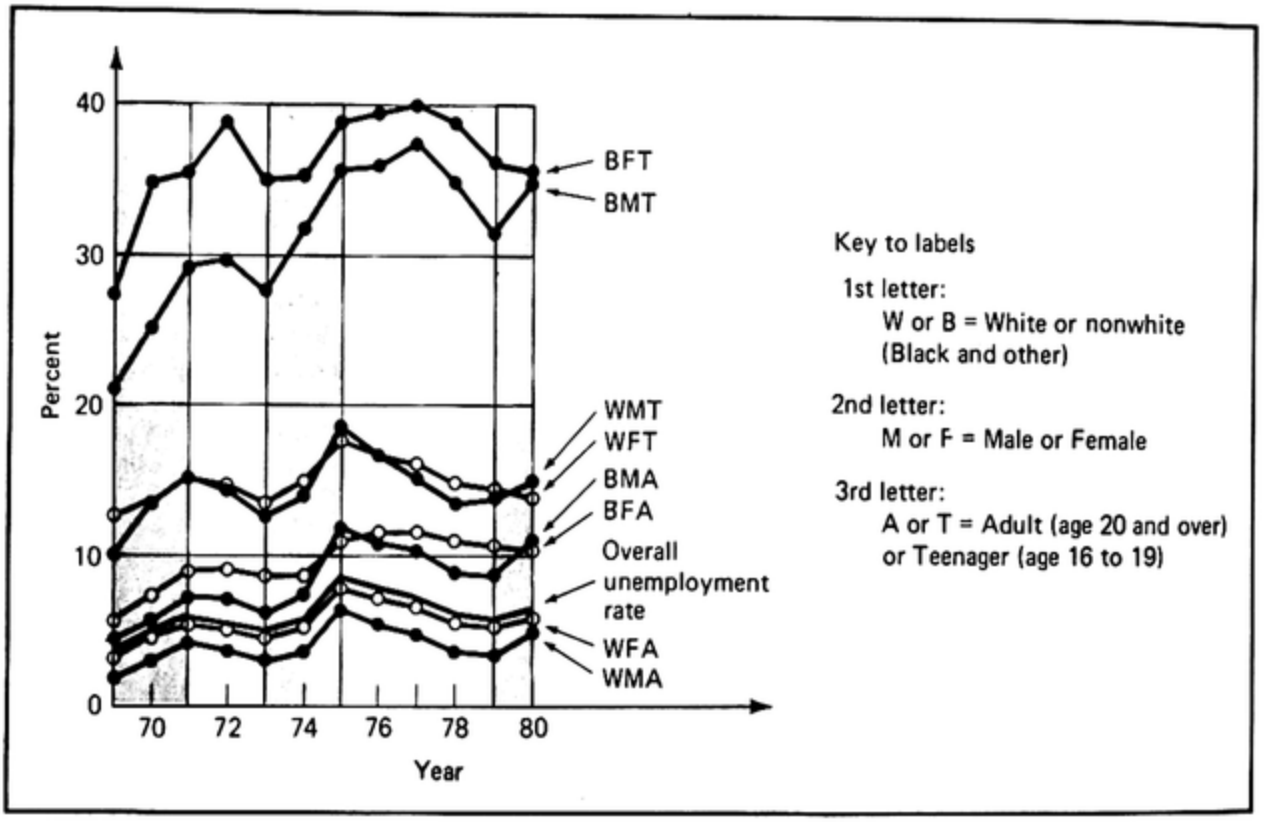


Figure 8 Unemployment rates for various race, sex, and age groups 1969–1980

Unemployment rates for nonwhites are very much higher than those for whites, those for teenagers are above those for adults, and those for females are usually above those for males. But, in general, the unemployment rates for various population groups rise and fall together over the business cycle.

Source: *Economic Report of the President*

find the overall unemployment rate from 1969 to 1980 toward the bottom of Figure 8. The business cycle reached a peak in 1969, a trough in 1971, a peak in 1973, and a trough in 1975. The alternate shades in the diagrams represent periods of contraction and expansion.

The thing to notice about Figure 7 is the substantial number of industries that individually follow the overall pattern, particularly the two manufacturing industries (2) and (5) and construction (7). These are the most cyclically vulnerable industries. In three other industries—(1), (4), and (6)—the overall cyclical declines show up just as a sectoral slowdown in growth during one or both of the recessionary periods. Only sectors (3) and (8) and the miscellaneous grouping (9) fail to reflect the general cyclical pattern. In the chapters that follow, we will try to show you why the cyclical pattern is shared throughout most of the economy. What really matters is not the *number* of industries conforming to the cyclical pattern, but their share of product. In 1969, the three industries that most closely conformed to the overall cycle produced 36 percent of the collective product, the three for which recession was just a slowdown in growth produced 42 percent, and the nonconforming industries produced 22 percent.

Unemployment

An unemployment portrait of the 1970s organized by geographic region would show much the same picture as the output series show. Those areas of the country in which employment is dominated by cyclically sensitive industries—particularly the industrial heartland around Chicago, Detroit, Cleveland, and Pittsburgh—are subject to violent fluctuations in hiring and firing. In these industrial centers, you can measure the unemployment rate by the amount of soot in the air. In commercial

centers like Boston, Atlanta, Denver, and San Francisco, where recessions are much milder, the fluctuations in unemployment are correspondingly mild.

Another perspective on unemployment comes from looking at the unemployment rates of various race, sex, and age groups: whites and nonwhites, males and females, adults and teenagers. Figure 8 shows these comparisons. Look first at the relative positions of the series. In nearly every case, the unemployment rates for nonwhites are above those for the corresponding groups of whites, for females above those for corresponding males, and for teenagers above those for corresponding adults. The unemployment rates among black, Puerto Rican, Chicano, Native American, and other nonwhite teenagers are staggering. The Great Depression lives on in the lives of these young nonwhite Americans. Under a more oppressive government, you might be arrested for incitement to riot just for publishing these figures.

As you try to appreciate how widespread the effects of the business cycle are, though, you should concentrate on the ups and downs, not on the relative positions. When you do this, you will see that the peaks and troughs of the various series coincide. The only exceptions are among young nonwhites, for whom good times sometimes just mean a slowdown in the rate of increase in unemployment.

Consumer prices

Recall from our earlier look at long-term trends in consumer prices that the inflation of the 1970s differed markedly from that of earlier periods in our history. In general, the periods of inflation have been periods of low unemployment. High unemployment has usually been accompanied by stable or falling prices. Not so the stagflation of the 1970s.

Figure 9 shows that the pattern of rising prices in the face of high unemployment was not confined to a few commodity groupings. The cost-of-living indexes presented in this diagram cover all of the major necessities in the consumer "market basket." Every single category of prices shows a strong uptrend throughout the period, with food and medical care leading the way. Particularly disconcerting is that without the shading that marks recessionary periods, these diagrams would not tell us which years were bad and which were not. Prices went up through both prosperity and recession.

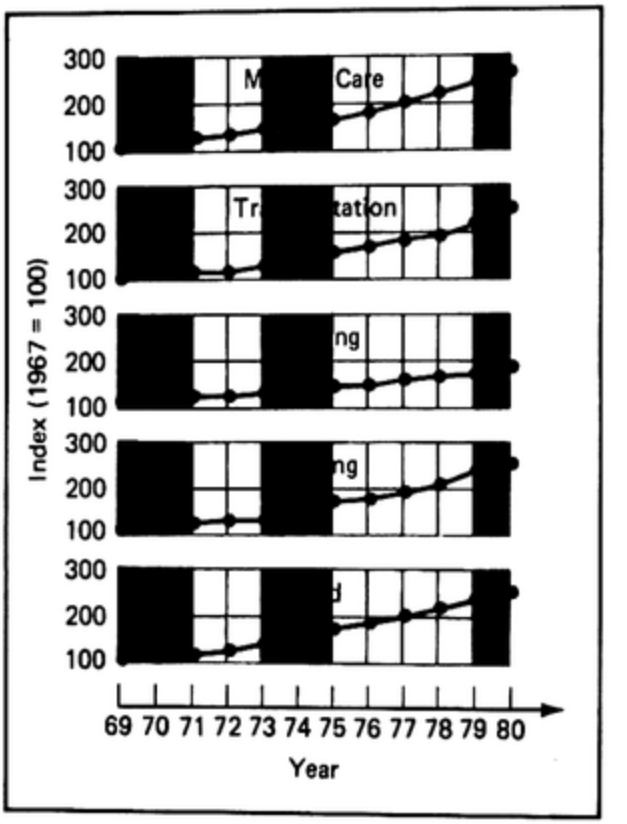


Figure 9 Selected consumer price indexes 1969–1980

During the 1970s, prices of all major categories of consumption goods and services rose steadily regardless of whether times were prosperous or depressed.

Source: *Economic Report of the President.*

Common stock prices

Stock prices have a fascination of their own, particularly for speculators. But except for periods of widespread speculation, the swings in stock prices are mainly interesting to economists for what they tell about the state of investor confidence. Re-

member from our earlier discussion that the stock market was largely stagnant during the 1970s after 20 years of an almost unbroken uptrend in prices. Much of this stagnation has to be attributed to the succession of crises that rocked public confidence, particularly the combination of high unemployment and rapid inflation. Figure 10 details some of the market fluctuations during the 1970s. Each of the series shown on the graph represents one of the groups of stocks traded on the New York Stock Exchange. As you can see, the cyclical swings in stock prices mirror the business cycle, declining during recessions and rising during periods of general business prosperity. In fact, the stock market tended to lead the swings in the unemployment rate, turning up in 1971 and down in 1973 in anticipation of the swings in production and employment. This is what you would expect of an indicator of expectations. The stock market would be a good predictor of things to come if it didn't also tend to predict recessions that never happen.

For someone trying to learn how the economy fits together, probably the most interesting feature of Figure 10 is the degree of similarity among the patterns of movement in the prices of quite different groups of stocks. The reason for this is simple. The participants in the stock market—the individual and institutional investors whose expectations are mirrored in stock prices—understand quite well what we have been trying to establish by looking at the cyclical record of the 1970s: the interdependence of various sectors of the economy and how their fortunes rise and fall together.

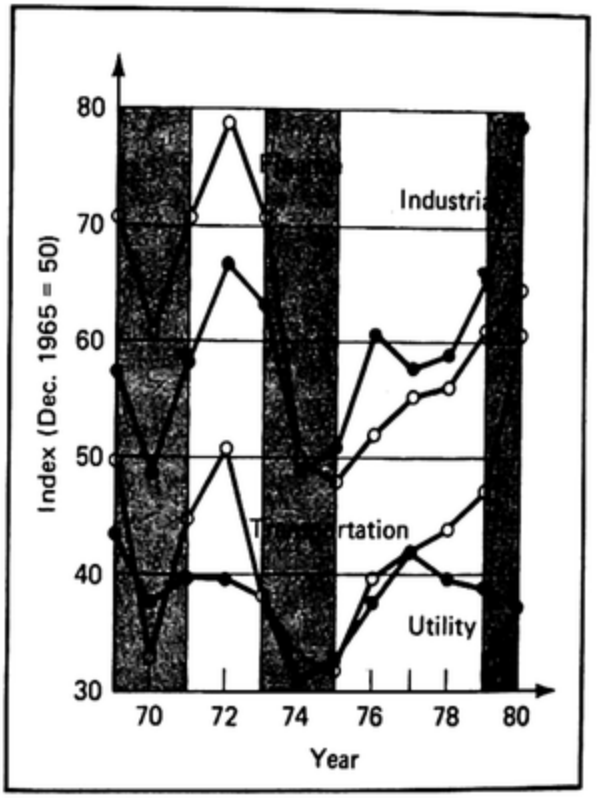


Figure 10 Common stock prices of selected groups of stocks traded on the New York Stock Exchange 1969–1980

Changes in stock prices usually lead changes in the overall business cycle, although this pattern is not perfectly reliable. But the prices of major groups of stocks can be counted on to rise and fall together.

Source: *Economic Report of the President.*

Macroeconomics and the 1970s

The 1970s were just awful. Some people stopped reading the newspapers because there was never any good news. Others took solace in the fact that although things always seemed to be getting worse, sometimes they got worse less fast than they did at other times. Unemployment was high, inflation was high, productivity gains were low, real wages stopped rising for the first time since anyone could remember, stock prices were depressed, and the government budget leaked a torrent of red ink. Economists who for so many years had offered smug advice on everything from the balance of payments to the heroin epidemic seemed strangely quiet. They are still quiet, or at least less brassy than they used to be. The Keynesians are the most puzzled, because their traditional remedies seemed particularly ineffectual in the 1970s. The monetarists have a lot to say

that makes sense, but their policies have never been put to the test. The Marxists say "I told you so," but since they are always predicting collapse, it is hard to see that the collapse of the 1970s was a stunning verification of their theory.

This is, in fact, a good time for economists to be modest, for they have a lot to be modest about. The following chapters will help you understand the 1970s and a whole lot more. They won't give you certain answers, but they will suggest possibilities. What you can hope to gain by working through them is an understanding of macroeconomics, its strengths and its failings. If they prepare you to think through the macroeconomic problems of your own day in an analytical, thoughtful way, they will have served their purpose. And someday your picture may appear in a later edition of this book along with those of Marx, Keynes, and Friedman.

Summary

This chapter has been mainly historical, looking at the long sweep of American economic development from the Civil War to the present. It has also looked closely at economic fluctuations during the decade of the 1970s. As you begin your study of macroeconomics in the coming chapters, there are several features of this historical survey that you should remember.

1. Over the past 100 years, the nation's output of market goods and services (GNP) has grown at a rate between 3 and 4 percent a year. Since population growth over the same period was less than 2 percent a year, much of this output growth represented an increase in per capita goods and services.
2. This growth was by no means steady, but consisted of alternating periods of prosperity and depression.
3. Fluctuations in output are mirrored in fluctuations in the unemployment rate, the fraction of the labor force that cannot find work. This rate has ranged from a low of 1½ percent during World Wars I and II to a high of 25 percent in 1933, the worst year of the decade-long collapse known as the Great Depression.
4. Although prices have risen considerably on average over the 120 years since the beginning of the Civil War, substantial periods of price decline can be found. The general pattern up until the end of World War II was for prices to rise in wartime and to fall in peacetime. Since World War II, prices have risen most of the time. There has been more inflation in the past 30 years than in the previous 90.
5. One of the most significant trends in the economy over the past 50 years has been the great growth in the relative importance of the federal government. The share of GNP going to the federal government was about three times as large in 1980 as it was in 1929.

Key concepts

Macroeconomics
Gross national product
The business cycle
The Great Depression
The unemployment rate
Consumer price index
Common stock

Questions for review

1. Suppose that you examine GNP data from two different countries. In one, GNP in constant dollars has fallen during the past two years. In the other

country, real output has risen dramatically over the same two-year period. Can you therefore assume that the second country has a stronger economic growth trend than the first? Explain.

2. A comparison of Figure 2 and Figure 3 shows that there has often been an inverse relation between price changes and changes in unemployment. Jot down some of your own ideas on why this inverse relation might hold. This issue will be presented in detail in later chapters. When you are studying those chapters, check your original ideas against the material in the text. How correct or complete were your initial impressions?
3. Figure 7 highlights the industries that are most vulnerable and least vulnerable to cyclical fluctuations. Explain why construction and durable goods (cars, large appliances) might be so sensitive to fluctuations in GNP, while finance and agriculture are less sensitive.
4. In your own words, explain why common stock prices tend to rise and fall together, as Figure 10 shows.

23

National Product and Income

As you read and study this chapter, you will learn:

- how to visualize the structure of the economy with a diagram of the circular flow of spending, production, and income
- how to measure national output in ways that give insight into the circular flow
- why some sectors of the economy must run deficits if others are running surpluses
- how price and output trends are separated with the use of price indexes

If you have ever flown over New York City on a clear day, you will never forget what it looks like. Its landscape is an unmistakable mosaic of islands and rivers. Even more arresting than its physical geography, though, is the evidence it gives of great economic forces at work. Its buildings seem to dwarf the Rockies, and its bridges outnumber those that cross some of the world's greatest rivers. Even its graveyards seem bigger than life. At the turn of the century, H. G. Wells, the science fiction writer, wrote that New York will someday make a fascinating ruin. Maybe so. But for now, it is a marvelously vibrant center of human activity.

Most of us live somewhere else. Our cities and towns are grayer and flatter. We might have a hard time telling someone how to recognize them from the air. Yet their *collective* economic might surely overwhelms that of New York.

Think about trying to measure the economic vitality of such a city from the air. It is possible in these days of aerial spying to take photographs from the stratosphere that show more visual

detail than anyone would ever be interested in. Would a detailed aerial photograph help you to grasp the economic life of a city? Probably not. If it were blown up large enough to show what people were doing, not only would there be too much detail to comprehend, but the photomap would have to be nearly as large as the city itself.

A far better charting of economic life could be read from an infrared photo, which detects thermal energy. It shows where a city is hot and where it is cool. Concentrations of production and commerce show up as hot spots. Highways and railroads show up as lines tracing the links between the hot spots. The graveyards don't appear at all. Such a photo, taken with a wide-angle lens, gives a good overall picture of where the economic action is and where the links between the centers of this activity are.

If you try to picture how the economy of an entire country fits together, you will see right away that even the infrared photograph is of little use. First, geography is not always the most interesting dimension of the layout of the national economy. Dividing the economy into industries is usually more revealing than dividing it into regions. Second, heat isn't a terribly good indicator of economic activity on a national scale. Death Valley is plenty hot even in winter, but there is much more going on in Minneapolis, where it is bitterly cold. Third, to show enough detail, the map would have to be too big to take in, in one scrutiny.

To understand how an economy works, you need an analytical, not a geographical, map. This map should separate the economy into sectors according to how they *function* in the economy, their role as an ongoing, changing pattern of human activity. It should stress the links among these functional sectors and highlight how they interact with one another to produce this pattern. Finally, it should be quantitative enough to include comfortably the rich variety of documented economic experience. To provide yourself with a blueprint or road map of the U.S. economy, you will learn to use a system of *national accounts,* which is simply a conceptual framework that puts together data from all the diverse parts of the economy. Developing such a system of accounts requires abstraction, the ability to decide what to include and what to leave out. This abstraction is obviously necessary from a practical viewpoint. But even if you could include everything, you wouldn't want to because too much detail would hide the relationships you are trying to highlight.

The accounting system generally used in market capitalist countries such as ours treats all market-oriented activities from which people legally make a living as productive. Nearly all nonmarket activities are treated as unproductive. This means leaving out nearly all home production, which must take up about half of our working hours. Raising someone else's children for pay is a "productive" activity; raising your own children is not. This division between productive and unproductive work is partly a practical matter. It is difficult to measure the worth of goods and services for which no market transaction is recorded. But this way of viewing productive activity is also a natural extension of the attitudes of many participants in the market system. As President Coolidge said, "The business of America is business." If you agree with this division of human activity into what is productive and what is not, you have a lot of company. The governments of all the market capitalist countries do their national accounts this way. Many economists, however, are uncomfortable with this way of defining productive activity, and some have developed alternative ways of analyzing the economy. But the provision of data on how the economy

is performing is very largely a government monopoly. Since it is almost impossible to talk intelligently about economic life without data, we—and the news media that package economic data for public consumption—are largely stuck with the economic maps provided by government agencies. Since you will very likely have to rely on these media for news of economic events, you should learn to organize your thinking around the concepts used in the official U.S. national accounts.

The circular flow of goods and services

In presenting something so complex as a national economic map, it is hard to know whether to start with a description of the parts and then trace the interconnections, or go the other way around. The best way seems to be to alternate: first, give a quick description, then trace the main connections, then refine the description, then trace more connections, and so forth. This is probably how you were taught a foreign language. First you learned a few words and some basic grammar. Then you put them together in simple sentences. Next you learned more words and more complex grammar; then you combined these with what you had learned earlier to produce more complex sentences.

A two-sector economy

Imagine an economy without taxes, government spending, or foreign trade. Its economic landscape consists of only two distinct features. There are *firms*, which are producing units. Their products consist of *goods* (like cars, oatmeal, and books) and *services* (like manicures, legal advice, and concerts). For simplicity, it is easiest to use the word "goods" to refer to both goods and services. There are also *households*—families, others living together, and people living alone. The households buy goods from the firms. Most of the households have one or more members working in the firms, supplying *labor services* to the production process. Some households own the firms, either directly or through ownership of stock. Others are lenders, providing funds in exchange for interest. In a legal sense, these owners and lenders provide the natural resources, buildings, equipment, and inventories needed by the firms. They are said to provide *capital services*. ***Together, labor and capital services are called*** **factor services.**

All these households and firms are tied together in a network of mutual dependence. The firms need customers and workers. Under our system of private ownership of the means of production, they also need people who are willing to supply capital. The households need consumption goods supplied by firms. They must also have some way of paying for them, so that they also need jobs for which they get *wages*, or they must own capital from which they get some form of *property income*, such as rent, interest, or profit. ***Wages and property income make up the*** **factor income** ***from which households get their ability to buy goods and services.***

This mutual dependence between firms and households is illustrated in the bottom loop of Figure 1, which pictures what you know as the *circular flow* of goods and services. In this figure, the entire collection of households is represented by the block called the ***household sector.*** All firms are represented by the block called the ***business sector.*** Some of the *outputs* of the business sector go to the households in the form of consumption goods. Factor services supplied by the household sector make up part of the *inputs* into the business sector.

In any modern economy, there is a second mutual dependence, involving rela-

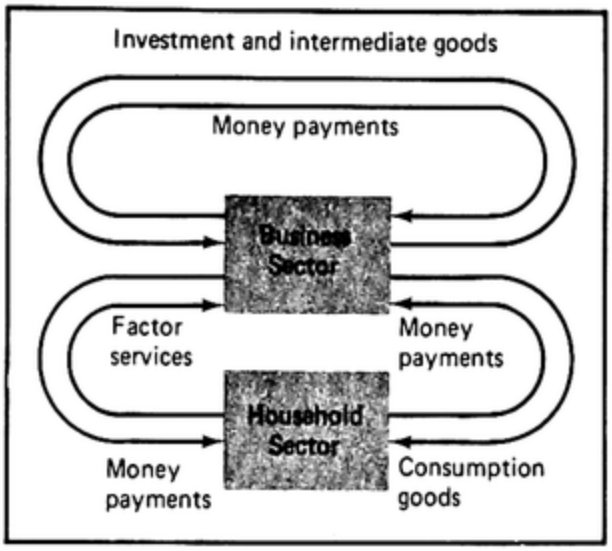


Figure 1 The circular flow in a two-sector economy

This simple diagram of the circular flow shows the two mutual dependencies in an economy made up of households and firms. Investment and intermediate goods pass from one firm to another. Consumption goods pass from firms to households. Households supply factor services to firms. Money payments flow in opposite directions from the flow of goods.

tions among firms. Complex production processes such as making steel involve many stages: iron mining, coke reduction, smelting, refining, rolling, stamping, finishing, shipping, and selling. Many of these stages are performed by different firms, or different divisions of the same firm. They produce goods and services for one another. Some of these, in finished form, represent additions to the stock of durable productive facilities: buildings, machinery, equipment, company parking lots, mine shafts, and many other kinds of long-lasting additions to productive capacity. These are called ***investment goods.*** Others represent goods and services that will enter into further production and be directly used up in the process: flour, steel, the services of corporate law firms, brazing rods, typing paper, and the like. These are called ***intermediate goods.*** The flow of investment and intermediate goods from firm to firm is shown in the top loop of Figure 1. Since this flow feeds back into the production process, it comes from the business sector as an output and reenters as an input.

All these goods and services have to be paid for, of course. Each of the loops in Figure 1 is a two-way street. Goods and factor services go one way, money payments go the other. Some firms pay other firms for investment and intermediate goods. All firms pay incomes to households in exchange for the labor and capital services the households provide. Households must pay for the goods and services they consume. And like you, they certainly expect to be paid for their work.

This exchange of products and payments is largely responsible for one of the important facts we discussed in the last chapter: the tendency for most of the sectors of the economy to fluctuate together. If households cannot sell labor services, they cannot afford consumption goods. If the firms making consumer goods can't sell them, they can't buy investment or intermediate goods. And they can't employ workers or pay dividends. Through these channels, bad times spread from one part of the economy to another. Prosperity spreads by the same paths. Understanding these paths is essential to understanding how the economy as a whole grows and fluctuates.

A four-sector economy

Once you can envision the general outlines of an economic system made up of private firms and households, it is easy to extend your vision to include foreign trade, the *foreign sector* of the economy, and government, the *government sector.* Let us begin with trade, since it is the easier of the two.

Most exports that the U.S. economy sells to other countries are the products of private business, much like goods and services that are sold domestically. The citizens of one country also export factor services to the firms of another country. If an American engineer works for a Canadian mining company, the United States is exporting labor services to Canada, no mat-

ter which country the engineer lives in. If an American firm owns a European subsidiary, the owners of the American firm are exporting capital services to Europe. Thus, the total of exports is made up of both goods and factor services.

Most imports that the U.S. economy gets from other countries are intermediate goods. True, we import many identifiably foreign consumer goods, such as Hong Kong suits, French wines, and Japanese cars. But even these are intermediate goods in the sense that they pass through domestic retail outlets on their way to market. To the extent that foreign nationals work in or receive property income from American firms, we also import factor services.

Government is a little more complicated. The flow of goods and services from the business and household sectors to federal, state, and local governments is simple enough. The government buys the labor services of firemen, policewomen, teachers, and bureaucrats, all from households. It also buys fuel, missiles, red tape, and a wide variety of other goods from the business sector. But what does it supply in return?

Your answer to this question will tell a lot about your politics. If you reply "a headache," you are probably a Libertarian. If you say something about the government's being "the executive committee of the capitalist class," you are a Marxist. If you cite a long string of specific government programs that help people in your locality, you are probably running for office as a Democratic or Republican incumbent. If you simply answer "the product of government," you are an employee of the U.S. Department of Commerce, which prepares the national accounts.

There is a fairly obvious list of things that governments provide: defense, education, police and fire protection, administration of civil law, highways and waterways, sanitation, and much more. Unlike the products of private firms, however, the products of government are not sold, so that there is no good way to determine the collective value that people place on them. True, we pay taxes to all levels of government. But there is an element of compulsion in taxes that is not there when you buy *The New York Times* or when a business firm buys a new copying machine. Just try refusing to pay your school taxes because you don't like the pictures in the Dick, Jane, Sally, and Spot readers your school district may still be using.

The Commerce Department simply values the product of government at what the government sector pays for its inputs, the goods and labor services it buys from firms and households. In effect, the government provides "governance" to the private sector at cost. Figure 2 conceives of the government in this way. Its lower loop shows goods and labor services flowing

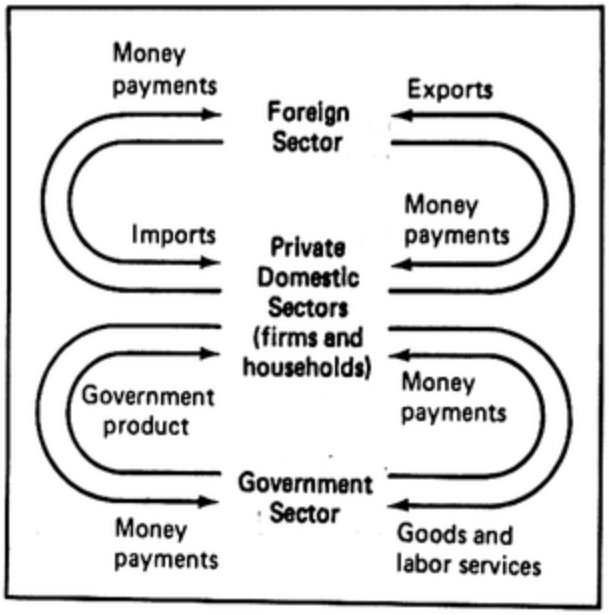


Figure 2 The circular flow in a four-sector economy

Besides the flows of goods and services shown in Figure 1, an economy with foreign trade and a government exchanges goods and services between its private domestic sectors and the foreign and government sectors. For each of the flows of goods and services, there is a corresponding money flow, but government product is usually valued at its cost to the government rather than at the cost of taxes paid by the private sectors.

from the private domestic sectors (firms and households) to the government. The government sector bestows government product on the private sectors, some directly to business, some to households, and some to both. Figure 2 also shows the connections between the private domestic sectors and the foreign sector. Exports flow from firms and households to the foreign sector. Imports come back. Corresponding money flows run in the opposite directions. The block in the middle contains the flows shown in Figure 1.

Measuring national output and income

If you can clearly envision the circular flow of goods and services, you have a very good understanding of how the sectors of the economy fit together. But that is not enough. You must also understand how the flow is measured. So much of economics is quantitative that you will miss a lot if you are not at ease with measurement. So look back at the circular flow diagrams and think about measuring what is going on there, remembering that the central block of Figure 2 contains within it the complications of Figure 1.

If the circular flow were really a circle, it would be possible to pick any point on it and then measure the *market value* of the goods and services flowing past that point during some *time period*, say a year. This would give a reading on the level of economic activity, measured in *dollars per year*. You can see from the diagrams that this won't work. Even the highly abstract diagram scheme made up of Figures 1 and 2 is not a circle, but a figure eight within a figure eight. To measure what is going on, you have to break into the flow in several places at the same time and combine the readings taken at the various points. For this to make sense, the points chosen must have something in common.

The gross national product

The usual measurement taken of the circular flow is called the *gross national product*, or GNP. Figure 1 in the last chapter showed you more than a century of growth and fluctuations in GNP. The GNP is the market value of the country's total output (over some time span), after the value of imports and of the intermediate goods used up in the production process have been subtracted. There are several ways to measure this magnitude, all of which give the same number. Three of the measures are particularly useful because of how they fit into the theory of what governs changes in GNP:

1. The value of goods delivered to ***final demand.***
2. The ***value added*** in the producing sectors.
3. The ***gross national income*** originating in the producing sectors.

Each of these is calculated by asking a different set of questions about the circular flow. Suppose you asked what consumers spent on goods and services last year, how much businesses invested in plant, equipment, and inventories, what goods and services governments bought, and what U.S. exports foreigners imported—all valued at market prices. If you added up the answers and subtracted the value of American imports (which are produced abroad), you would have the total value of finished goods delivered to the household, business, government, and foreign sectors last year. This is appealing as a measure of *national output*, since it counts only finished goods (not intermediate goods or services), and it subtracts that part of their value that reflects production in other countries. This

measure, called *deliveries to final demand*, is one way to measure GNP.

These goods that are delivered to final demand are produced someplace. Take a bottle of beer. It may cost 60 cents plus tax at the corner grocery. Part of this price pays for the services of the grocer who sells it to you and of the distributor who delivers it to the grocer. Part goes to the brewer and to the bottle manufacturer. Part goes to the farmers who grow the barley and the hops. Some pays overage, overweight football players to advertise it on TV. Each of these stages of its production and marketing contributes to the price of the beer. So does the sales tax collected by the state. Every good delivered to final demand has its value built up in a succession of similar stages. And every productive enterprise contributes its share to the value of some final product at some stage. If you asked each producer the value of its output last year, less the value of the intermediate goods it used up, you would have its contribution to goods that were destined for final demand, after they had passed through all their stages of production. You would have its *value added*. Even if these goods were sitting in someone's inventory at the end of the year, they would still be part of final demand, since investment in inventory is part of the business sector's final demand. Thus, the sum of value added by all firms combined is equal to the value of deliveries to final demand. This, too, is GNP, calculated in a second way.

Finally, suppose you asked a firm's accountants what happened to the income the firm got from its value added last year—the difference between the value of its outputs and the intermediate goods it used up in their production. The accountants would tell you that some of the income was paid to the government as sales, property, and other indirect taxes; some went to employee compensation (before taxes); and some became property income—rent, interest, and profit (all before taxes), plus depreciation allowances that compensate for wear and obsolescence of fixed capital. These would account for every penny of value added, and their total from all firms would equal the GNP. However, since it would be measured as the sum of incomes, it would be called the *gross national income* (GNY).

The following sections describe value added, GNY, and deliveries to final demand in more detail, and spell out how they are related to one another.

Input-output accounting

Occasionally, the Department of Commerce publishes detailed studies of the flow of intermediate goods from industry to industry. These studies, called ***input-output tables***, form the basis for piecing together current data into estimates of value added in the various industries. Since they are very expensive to prepare, they are done rather infrequently. The most recent table available covers the year 1972. Although it is now out of date, it is the best available source of insight into the relationships among inputs, outputs, and value added in the American economy.

Input-output analysis has a usefulness that extends far beyond its contribution to national income accounting. Because it provides a quantitative picture of the relationships among the various producing sectors of the economy, it shows the paths by which output and price changes in one sector influence outputs and prices in all the others. You will encounter some of these applications in later chapters. But for now, we will focus on how input-output data contribute to the measurement of national output.

Figure 3 shows the logical structure of an idealized input-output table, describing the annual flows of goods in an economy in which every industry produces only one

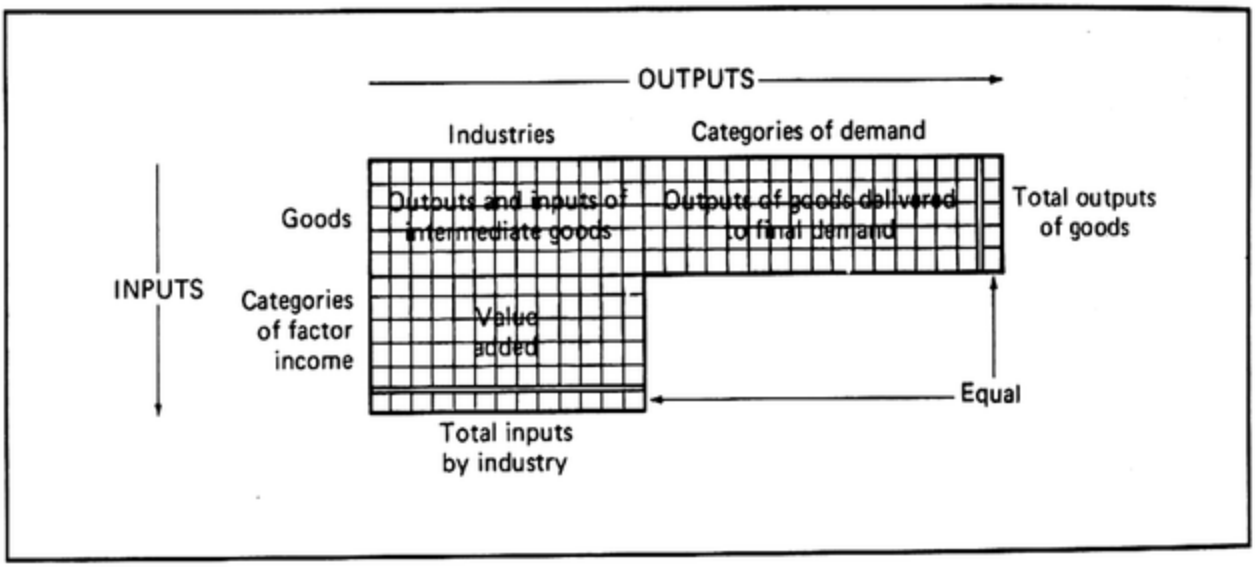


Figure 3 The logical structure of the input-output system

Understanding the relationships among inputs and outputs is the key to understanding the equivalence of the various ways of looking at GNP. Reading down the figure, you can find inputs of intermediate goods into various industries plus value added in those industries. Reading across, you can see where the various goods produced by these industries were used. GNP is equal to the sum of value added by all industries or the sum of deliveries to final demand of all goods. It is smaller than the sum of total inputs to industries or total outputs of goods, since these totals include intermediate goods produced and used up.

category of goods. It is divided into several blocks, each of which is divided into rows and columns. The rows tell what happens to the outputs of goods. The columns tell the composition of the inputs into an industry. Both inputs and outputs are flows measured over a year.

Consider a typical row, corresponding to the output of livestock, say. The first set of entries in that row shows the dollar value of livestock used as intermediate inputs by other industries, such as meat packing. The second set of entries shows the value of livestock delivered to final demand: consumption, investment, government purchases, and exports, less the value of livestock imports. Included in investment are any net additions to the inventory of livestock over the year. At the extreme right is a row total. Since additions to or subtractions from the inventory of livestock are included in final demand, the sum of intermediate and final uses accounts for the entire output of livestock, and the row total equals the value of this output.

Now consider the livestock column, whose position in the sequence of the columns is the same as livestock's position in the sequence of rows. The first set of entries shows the values of various intermediate goods used up by the livestock industry. The second set of entries is made up of the various factor incomes paid by the industry. The sum of these factor payments, as you know, is equal to value added. At the bottom of the column is the sum of the items above it, the total value of intermediate goods used up plus value added. This must equal the total value of livestock production. Thus, the entry at the bottom of the livestock column equals the entry at the extreme right of the livestock row.

As you know, economy-wide value added must equal total deliveries to final demand. But must this also be true on an industry-by-industry basis? Not necessarily. Suppose industry A uses 100 interme-

diate goods and has 100 in value added. Its total inputs add up to 200, and its total output is valued at 200. If it delivers 150 of this output to other industries as intermediate goods and 50 to final demand, its value added is larger than its deliveries to final demand. If it delivers 50 of its output to other industries and 150 to final demand, its value added is smaller than its deliveries to final demand. Only if the uses of its output are split 100–100 will value added equal deliveries to final demand. Thus, it would be mere coincidence if an industry's deliveries to final demand were equal to value added. ***This equality holds for all industries combined, of course, precisely because the intermediate goods they jointly supply are identical to those they jointly use up.***

A real input-output table is somewhat messier than the idealized description of Figure 3, mainly because there is not a perfect correspondence between industries and outputs. This stems from the fact that so many industries produce by-products that are different from their main products. The U.S. input-output tables are large, as well as messy. Table 1 lists the 79 industries included in the 1972 study. If you want to look at the flows of goods among these industries, you will have to consult the April 1979 issue of the *Survey of Current Business*, put out by the Department of Commerce. But Figure 4 illustrates some of its main features with data from the first two industries, (1) livestock and livestock products and (2) other agricultural products.

Value added and gross national income

Look at the first two columns of Figure 4. They show the total outputs of the two industries. Livestock production was $43.339 billion and other agricultural production was $35.080 billion. (You can find these figures at the bottoms of their respective columns.) But both industries used intermediate inputs. In fact, the total intermediate inputs into the livestock industry were worth $33.776 billion, accounting for more than three fourths of the value of its output. *Value added* in this industry was therefore only $9.563 billion. You can see from the top of column 1 that most of the intermediate inputs into livestock raising came either from industry 1 itself (as animals shipped from one part of the industry to another for fattening) or from other agricultural producers (as raw animal feed). The other intermediate inputs, which are not shown to save space, came mainly from industry 14, food and kindred products (as processed animal feed).

Value added, as you know, is the difference between the value of an industry's output and the value of the intermediate goods it uses up. Since this difference is the source of the wages, property incomes, and indirect taxes paid by the industry, it equals the *gross national income* originating in the industry. Those portions of gross national income originating in industries 1 and 2 were $9.563 billion and $19.600 billion, respectively, just equal to the corresponding value added totals. In the livestock industry, $1.854 billion was *compensation of employees* (wages, salaries, and supplements), $0.794 billion was *indirect taxes* (mostly property taxes on land and buildings), and $6.914 billion was *property income* (interest, rent, depreciation, and profit). The division of value added in the other agricultural products industry was roughly similar. The income in these two industries is disproportionately weighted toward property income, however, since land rent is such a major cost item in agriculture. In industry 14 (food and kindred products, which is a manufacturing industry), about 60 percent of value added was employee compensation. This is more representative of the economy as a whole.

Table 1 ***Industries included in the 1972 input-output study***

1	Livestock and livestock products
2	Other agricultural products
3	Forestry and fishery products
4	Agricultural, forestry, and fishery services
5	Iron and ferroalloy ores mining
6	Nonferrous metal ores mining
7	Coal mining
8	Crude petroleum and natural gas
9	Stone and clay mining and quarrying
10	Chemical and fertilizer mineral mining
11	New construction
12	Maintenance and repair construction
13	Ordnance and accessories
14	Food and kindred products
15	Tobacco manufactures
16	Broad and narrow fabrics, yarn and thread mills
17	Miscellaneous textile goods and floor coverings
18	Apparel
19	Miscellaneous fabricated textile products
20	Lumber and wood products, except containers
21	Wood containers
22	Household furniture
23	Other furniture and fixtures
24	Paper and allied products, except containers
25	Paperboard containers and boxes
26	Printing and publishing
27	Chemicals and selected chemical products
28	Plastics and synthetic materials
29	Drugs, cleaning and toilet preparations
30	Paints and allied products
31	Petroleum refining and related industries
32	Rubber and miscellaneous plastics products
33	Leather tanning and finishing
34	Footwear and other leather products
35	Glass and glass products
36	Stone and clay products
37	Primary iron and steel manufacturing
38	Primary nonferrous metals manufacturing
39	Metal containers
40	Heating, plumbing, and structural metal products
41	Screw machine products and stampings
42	Other fabricated metal products
43	Engines and turbines
44	Farm and garden machinery
45	Construction and mining machinery
46	Materials handling machinery and equipment
47	Metalworking machinery and equipment
48	Special industry machinery and equipment
49	General industrial machinery and equipment
50	Miscellaneous machinery, except electrical
51	Office, computing, and accounting machines
52	Service industry machines
53	Electric industrial equipment and apparatus
54	Household appliances
55	Electric lighting and wiring equipment
56	Radio, TV, and communication equipment
57	Electronic components and accessories
58	Misc. electrical machinery and supplies
59	Motor vehicles and equipment
60	Aircraft and parts
61	Other transportation equipment
62	Scientific and controlling instruments
63	Optical, ophthalmic, and photographic equipment
64	Miscellaneous manufacturing
65	Transportation and warehousing
66	Communications, except radio and TV
67	Radio and TV broadcasting
68	Electric, gas, water, and sanitary services
69	Wholesale and retail trade
70	Finance and insurance
71	Real estate and rental
72	Hotels; personal and repair services exc. auto
73	Business services
74	Eating and drinking places
75	Automobile repair and services
76	Amusements
77	Medical, educ. services and nonprofit org.
78	Federal Government enterprises
79	State and local government enterprises

Source: U.S. Department of Commerce, *Survey of Current Business*, April 1979.

Deliveries to final demand

The first two rows of Figure 4 show what happened to the outputs of livestock and of other agricultural products. By far most ($38.882 billion) livestock products were used as intermediate goods. Nearly all of this went to industry 14. Of the $1.822 billion output that went to final demand, most ($1.454 billion) was consumed. But of the output of other agricultural products, nearly half of the deliveries to final demand ($4.763 billion out of $9.973 billion) were exported.

At the bottom of the column labeled "Total final demand" you will find the figure $1,182.766 billion—a little over $1 trillion. This was GNP in 1972. It is the sum of the dollar values of all the various goods delivered to final demand as consumption, investment, government, and export goods, minus the values of these goods that were imported. At the extreme right of the table, a few rows from the bottom, you can find the same figure, calculated as the sum of value added in all the industries. Below it are the corresponding sums of the em-

Figure 4 The 1972 input-output study: dollar values (in $ millions)

This figure shows a greatly abridged version of the dollar value table from the 1972 input-output table. It enables you to see what happened to the output of livestock and livestock products and other agricultural products. It also shows for these two industries the importance of intermediate inputs and value added in the value of total output.

Commodity number	For the distribution of output of a commodity, read the row for that commodity. For the composition of inputs to an industry, read the column for that industry	Livestock and livestock products	Other agricultural products	Total intermediate use	Personal consumption expenditures	Gross private fixed capital formation	Change in business inventories	Exports	Imports	Federal government purchases: Total	Federal government purchases: National defense	Federal government purchases: Nondefense	State and local government purchases: Total	State and local government purchases: Education	State and local government purchases: Other	Total final demand	Total commodity output
	Industry number	1	2		91	92	93	94	95		96	97		98	99		
1	Livestock and livestock products	11,316	870	38,882	1,454	-----	439	111	−259	4	(*)	4	71	42	30	1,822	40,704
2	Other agricultural products	10,088	1,129	24,190	4,580	-----	2,053	4,763	−630	−982	(*)	−982	189	102	87	9,973	34,163
	← 89 Other Industries and Accounting Categories →																
I	Total intermediate inputs	33,776	15,480	----	-----	-----	----	----	-----	-----	----	----	-----	----	----	------	------
VA	Value added	9,563	19,600	----	-----	-----	----	----	-----	-----	----	----	-----	----	----	------	1,182,766
EC	Compensation of employees	1,854	2,556	----	-----	-----	----	----	-----	-----	----	----	-----	----	----	------	717,663
IBT	Indirect business taxes	794	701	----	-----	-----	----	----	-----	-----	----	----	-----	----	----	------	110,981
PTI	Property income	6,914	16,343	----	-----	-----	----	----	-----	-----	----	----	-----	----	----	------	354,122
T	Total	43,339	35,080	----	738,072	181,931	10,350	72,794	−76,199	102,126	73,513	28,613	150,693	63,816	86,877	1,182,766	------

Source: U.S. Department of Commerce, *Survey of Current Business*, April 1979.

ployee compensation, indirect taxes, and property income that make up gross national income. Both value added and gross national income, of course, must necessarily equal the value of deliveries to final demand of all industries combined.

Components of final demand To follow the theory of what determines GNP, you should both understand the components of final demand and have a compact notation for them. Picture the equality between GNP and final demand as

$$GNP = C + I + G + X - M$$

where

C = final output of consumption goods and services, or *consumption;*

I = final output of investment goods, or *investment;*

G = final output of government goods and services, or *government purchases;*

X = final output of export goods and services, or *exports;*

M = *imports* of goods and services.

These symbols will be used frequently, especially in the next two chapters.

Exports and imports, of course, are just goods and services sold to and purchased from other countries. Consumption, investment, and government purchases may seem like obvious activities, but the words mean different things to the economist and national accountant than they do in ordinary speech, so that it pays to spend some time studying them.

Start with *investment.* Investment consists of purchases of three different kinds of goods. The first is new housing and improvements on the stock of existing housing. The second is the purchase of plant and equipment by business firms, which is what most people probably mean by business investment. The third is by far the trickiest—net additions to business inventories.

All firms must hold inventories of raw materials, goods in process, and finished goods in order to carry on business. Most inventory holding is *deliberate* and *desired,* since business cannot operate without inventories. Try to imagine operating a grocery store without goods on the shelf or an assembly line with no partly finished automobiles and you will understand why inventories are essential.

However, some changes in inventory may be *unplanned* or *undesired.* If a firm sells less than it expected to, the unsold goods must be added to inventory. In a sense, a firm is forced to buy unsold goods from itself. Even this unplanned or undesired inventory change *is counted* as investment. Knowing this may help you understand why final demand and production are always equal. It's as though a quarterback forced "to eat the ball" were credited with completing a pass to himself. When it comes time to explain the theory of what determines GNP, however, it will be helpful to distinguish between planned and unplanned demand for inventories.

In the aggregate, ***inventory investment*** can be positive or negative, depending on whether the total *stock* of *inventories* is growing or declining. You may wonder whether this also applies to investment in housing and business plant and equipment.

In some situations, it is useful to think about ***gross fixed investment,*** the total of fixed capital goods that are produced. For analyzing patterns of changing employment in the machinery and construction industries, this is the relevant concept. But for measuring growth in industrial capacity and the housing stock, ***net fixed investment*** is what you are interested in. The difference between gross and net fixed investment is ***depreciation,*** which measures the rate at which capital is being lost

through wear, breakage, and obsolescence. A large part of investment replaces this loss, and simply maintains the capital stock rather than adding to it. If gross investment is bigger than depreciation, however, the difference makes a net addition to capital, and the stock of capital grows.

The investment figure included in GNP is made up of gross investment in housing, plant, and equipment, plus net additions to inventory. There is a corresponding national accounting concept that includes estimates of net investment in fixed or durable capital rather than gross investment. This is known as ***net national product,*** or NNP. Since the government's estimates of depreciation are subject to considerable measurement error, NNP is less often cited than GNP.

Consumption poses fewer problems of understanding than investment does. All household purchases of newly produced goods and services are included in consumption, except for the purchase of new housing, which is part of investment. However, consumption includes an estimate of the value of the services of houses occupied by their owners, who do not, of course, pay themselves rent. This is included along with market rental payments in the rent component of consumption. Leaving out the "implicit" value of the services of owner-occupied housing would seriously understate the consumption of housing.

Government purchases are also straightforward, if you distinguish between ***purchases*** and ***transfer payments.*** Purchases, which are part of final demand, include only expenditures on wages and salaries of government employees and on goods and services bought from the private sector. Transfer payments, which are not part of final demand, include Social Security, unemployment compensation, interest on the government debt, and many other kinds of payments. What distinguishes purchases from transfers is that purchases buy goods or labor services and transfers do not. Transfers, of course, provide incomes to the people who get them. But these people do not have to provide any goods or services in return. Transfer payments do not purchase inputs used in producing government product.

Sectoral surpluses and deficits

You may be wondering by now why you have to bother learning three ways to measure the same thing. There is an old saying that "one good reason is better than two." Isn't it also better than three?

The full answer to this question will gradually become apparent as you work through the following chapters. But this section of the present chapter should help you to see why all three ways of measuring the circular flow are important to understand what determines GNP. To keep your mind focused on the importance of the equivalence among final demand, value added, and GNY, think of it in the following way:

Final demand ⟶ Production ⟶ Income.

Except when resources, labor, and productive capacity are scarce, production always responds to changes in demand. Because of this, and because production always generates an equivalent flow of income (GNY), income also responds to changes in demand. But demands must also depend on income, at least in part. Although the circular flow isn't really a circle, its *causation* is circular. Seeing the dependence of income on demand is essential to understanding this circular causation.

Net taxes, business saving, and the distribution of GNY

Who do you think gets most of the gross national income—households, business, or government? In fact, about 70 percent of it

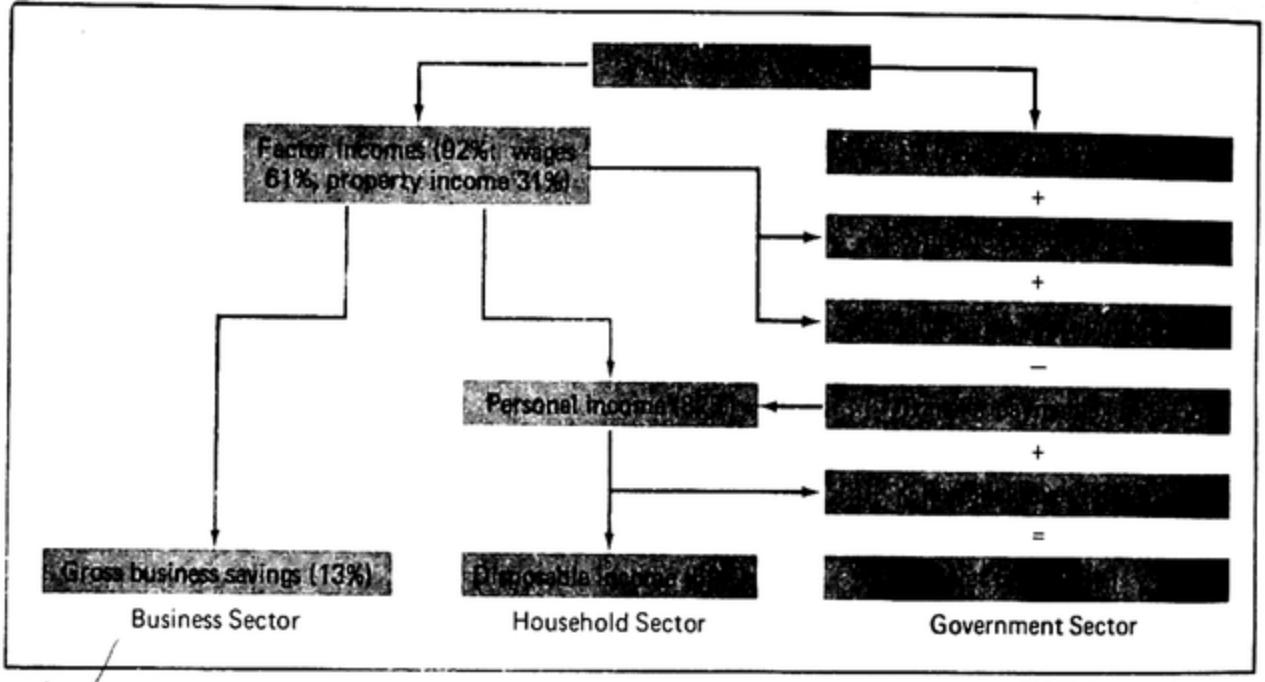


Figure 5 The sectoral distribution of gross national income (percentages are based on 1980 data)

This figure shows the main details of how gross national income was distributed among the domestic sectors of the economy during 1980. About 8 percent of value added was indirect taxes. The other 92 percent went to factor incomes, split 61-31 between wages and property income. Out of this property income, business saved 13 percent of GNY and paid 3 percent as corporate profits taxes. Out of the wages, 8 percent of GNY went into social insurance "contributions" (the largest portion of which is made up of payroll taxes for Social Security). This left 53 percent in the hands of households. Households also received transfer payments amounting to 14 percent of GNY. Thus *personal income* (before taxes) came to 82 percent of GNY: 15 percent from property income, 53 percent from wages, and 14 percent from transfer payments. Personal taxes took away 13 percent of GNY, leaving 69 percent as the *disposable income* of the household sector. Adding all the taxes and subtracting the transfers, the government sector collected 18 percent of GNY as *net taxes*. The other 13 percent was the gross saving of the business sector.

Source: *Economic Report of the President.*

usually goes to households, 17–20 percent goes to government, and 10–13 percent is retained by business. This distribution is important because demand is affected by the distribution of income among sectors. Income that goes to households directly influences demand. (You may know perfectly well how closely your spending tracks your income.) So, to a lesser extent, does income retained by business firms. Income that is collected as taxes has an even less direct and immediate impact on spending. Businesses and governments have great borrowing power, which they frequently use to finance deficits.

Since federal, state, and local governments collect nearly a third of GNY in taxes and distribute more than 10 percent of GNY in transfer payments to households, the tax and transfer system is one of the most important elements determining how GNY is distributed among sectors. Your understanding of the role of governments in the sectoral distribution will be a lot better if you see just where these taxes and transfers intervene in the circular flow. It will also help you to understand that business firms retain a bit more than 10 percent of GNY to maintain and expand the stock of capital facilities. This is known as ***gross business saving***, the sum of ***depreciation allowances*** and ***net business saving.*** The differences between taxes and transfer payments is called ***net taxes.*** The share of GNY that is distributed to households after tax is called ***disposable income.*** In 1980, gross business saving was 13 percent of GNY, net taxes were 18 percent, and disposable income was 69 percent. If you study Figure 5 and its accompanying legend, you will see some of the major features of this pattern of distribution.

Sectoral receipts and expenditures

These sectors of the economy that share in the proceeds of gross national income are, along with the foreign sector, those that purchase and pay for GNP. But final demand does not have to match up with income sector by sector. You must surely be aware of this. Households do a great deal of saving. Business firms have to raise a lot of outside money for expansion. The federal government usually runs a deficit. State and local governments usually run a surplus. Net exports are sometimes positive and sometimes negative. All of this is discussed in the newspapers and is part of your everyday experience.

The fact that GNP and GNY are necessarily equal restricts the surpluses and deficits of the various sectors. This is almost obvious once GNP written as a sum of final demands is set equal to GNY as a sum of sectoral incomes. In symbols:

$$GNP = GNY$$
$$C + I + G + X - M = YD + SB + TN$$

where C, I, G, and $X - M$ are consumption, investment, government purchases, and net exports. The new symbols in this equality are:

YD = disposable income;
SB = gross business saving;
TN = net taxes (i.e., taxes minus transfers).

All of these symbols will be used in later chapters.

Suppose that everything on the left-hand side is moved to the right and grouped according to sector. The result is:

$$0 = (YD - C) + (SB - I) + (TN - G) + (M - X)$$

household + business + government + foreign

Can you see that each of these four terms is the *surplus* of the corresponding sector—its net income minus its outlay? The difference between disposable (after-tax) income and consumption is obviously the surplus of the household sector, its saving. Gross business saving minus gross investment is the income retained by the business sector minus what it spends for expansion. This is usually negative, and the difference has to be made up by outside financing. Net taxes minus government purchases of goods and services is the government surplus—usually a negative figure for the federal government, whose budget is in deficit more often than not. The federal deficit is partly offset by a positive figure for state and local governments. Imports minus exports (the *negative* of net exports) is the surplus of the foreign sector. Our imports are a source of dollar income for the rest of the world. Our exports must be paid for in dollars. The difference is the surplus of dollars that foreigners get from trading with us.

Table 2 shows these surpluses in 1980, together with their component parts, expressed as percentages of GNP and GNY. Because of the equality between GNY and

Table 2 ***Sectoral surpluses and deficits (based on 1980 data)***

All figures are presented as percentages of gross national product and income*

Sector	Income	Expenditure	Surplus
Household $(YD - C)$	69	65	+4
Business $(SB - I)$	13	15	−2
Government $(TN - G)$	18	20	−2
Foreign $(M - X)$	13	13	0

Because of the equality of GNP and GNY, the surpluses of the four sectors must sum to zero.

Source: *Economic Report of the President.*

*Note that neither the income nor the expenditure column adds up to 100. To make them add up to 100, imports would have to be subtracted from both columns, since GNP (which is the denominator of all of these percentages) includes only net exports, not gross.

GNP, these surpluses add up to zero, although the individual sectors don't have to be in balance. As the equations in the previous paragraphs show, this must always be true. Any deficits must be balanced by positive surpluses somewhere else.

It is fairly easy to see why this is so if you understand that what appears as an expenditure in one sector must show up as a receipt in another. Suppose for a moment that within each sector of the economy receipts just equal expenditures, so that no individual sector shows either a surplus or a deficit. Now suppose that consumption increases, without an increase in income. Households now show a deficit, since expenditures are greater than income. The creation of a deficit in one sector must be balanced by a surplus somewhere else in the economy, but how this is done depends on how the increased consumption goods are provided. If firms meet the increased consumer demand by pulling goods out of inventory, then investment falls. Investment will now be less than retained earnings, and the business sector will show a surplus. Another possibility is that firms will increase production to provide additional goods. This will generate income that will show up as either retained earnings, creating a surplus for the business sector, or as additional income to households, bringing the household sector back into balance. Or the goods may be imported, so that a foreign surplus balances the household deficit. The important thing to see here is that in every case a change in the deficit or surplus of one sector of the economy must be balanced by a change in the deficits or surpluses of other sectors of the economy.

By itself, this is simply a fact of accounting that says nothing about how people behave. It is crucial, however, to the theory of how GNP is determined. To see why, suppose that households, domestic businesses, governments, and foreigners plan outlays of consumption, investment, and exports that do not equal their expected receipts, so that their surpluses and deficits do not add up to zero. The accounting relationships tell us that the results of these inconsistencies *must* be frustration: the plans and expectations of the sectors simply can't be met. Expectations must be revised and plans changed. Behavior must be altered. The result of the alteration of plans will spread frustration and surprise throughout the economy, causing widespread changes in production, income, and demand.

Measuring trends in prices and output

So far, you have been learning how economists measure output of GNP for one year. Often, however, it is important to look at what has been happening to national output over a longer period. Is it growing or declining? How fast? This introduces a challenging problem. Since GNP is measured by adding up many different goods and services, valued at their market prices, any change in GNP is made up of both changes in "real" or physical production and changes in prices. Anyone who cares about what has been happening to output over a long period must focus on the real changes. This involves separating the changes in output from the changes in prices.

Price trends are also important in their own right. The rate of price increase (inflation) is as much a key to the state of the economy and the lives of its people as is the rate of quantity increase. You know that from your own experience. Moreover, much of the income in the United States is *indexed*, that is, legally tied to the rate of price increase. For example, many labor contracts have cost-of-living clauses, and

Social Security payments are now tied to consumer prices.

To study variations in real GNP over time, the government has developed the ***GNP deflator*** to change GNP from current dollars to constant dollars (measured in the prices of a given year, called the base year). To study changes in prices, the U.S. Bureau of Labor Statistics has developed two measures: the ***consumer price index (CPI)*** and the ***producer price index (PPI).*** The CPI measures the retail cost of a typical urban family's consumption bundle, including goods and services, interest rates, and property taxes. The PPI focuses instead on interindustry transactions. It measures the cost of goods at intermediate stages of production, *before* the retail stage—raw materials, semifinished goods, and finished goods without their retail markups. Thus, the PPI is an important indicator of things to come. Most increases in producer prices will eventually affect retail prices and show up in the CPI.

Price Indexes

The CPI and the PPI are weighted price indexes. Each measures what happens over time to the total market cost of some fixed bundle of goods and services. The bundle that is "priced out" represents what was actually consumed or produced in a given year, called the **base period** ***or base year.***

The CPI, for example, charts what happens to the cost of some 250 major consumption goods and services. There is an overall index, and an index for each of several subgroups of purchases, such as food, clothing, and medical care. As you know from Figure 9 in the previous chapter, the prices of items in the subgroups usually change at different rates, even though they are moving in the same direction. Because there is no single rate of price change common to all 250 items, the rate of price change of the collective bundle is a weighted average of the rates of change of the separate prices. If the various rates of change are weighted by the *share* of the consumer budget spent on each in the base year, the result is mathematically equivalent to the rate of change of the price of the collective base-year bundle. For the CPI, the weights are currently based on an extensive survey of the consumption patterns of urban Americans during the 1972–1974 period. They will eventually be changed to keep abreast of changing consumer spending patterns, so that the index does not get out of date.

To understand how the CPI is constructed, think about a simplified example. Suppose that you represented a typical urban consumer in 1972, and that during that year you consumed only two goods, Toyotas and turkeys. Half of your total budget was spent on each. Now suppose that between 1972 and 1982, the price of Toyotas went up by a factor of 4, and the price of turkeys by a factor of 2. *On average,* therefore, prices rose for you by a factor of 3, or 300 percent. This is a 50-50 weighted average of 4 and 2. Since prices in 1982 would be 300 percent of prices in 1972, the price index in 1982 of your 1972 consumption bundle would be 300, relative to the base year of 1972 = 100.

Suppose, on the other hand, that you spent three fourths of your 1972 budget on Toyotas and only one fourth on turkeys. Assuming the same increases in the individual items, the increase this time would be a factor of 3.5, equal to $\frac{3}{4} \times 4 + \frac{1}{4} \times 2$. And the 1982 index would be 350, relative to the base year of 1972 = 100.

In principle, the official CPI is no more complicated than this example (although it is sometimes more controversial, as the box in this chapter indicates). Nor is the PPI. Each of them involves an enormous amount of information and calculation, but no real mystery. A particular base year's bundle of goods consumed or pro-

The Inflation of the 1970s and Early 1980s: How Much Was Statistical Illusion?

U.S. to Revise Housing Part of Price Index

Shift in '83 Could Trim Rise in Inflation Gauge

This headline appeared in *The New York Times* of October 27, 1981. Anyone who was inclined to follow up on such a humdrum leader could learn that the Bureau of Labor Statistics (BLS) planned to make certain technical changes in the CPI. Who would have thought this would appear at the top of page 1 of a national newspaper? During those troubled times, the 12 monthly entries in the inflationary fever chart were always page 1 material. But why was a technical change in the design of the thermometer, page 1 news?

The reason for the excitement lay in the second part of the headline—that the change would lower the measured rate of inflation, at least in times of rapidly rising housing costs. This would lower COLA (cost-of-living adjustment) payments built into wage contracts and Social Security payments. Thus, it was not simply a technical question. Whatever affects the distribution of income is always news.

During this period, the CPI was frequently suspected of exaggerating the rate of price increase when inflation was rampant. No one doubted there was inflation. The question was how much. And no one doubted the honesty of the BLS. The question was the suitability of its particular index to measure the cost of living.

One of the problems was the matter of weights. A "cost-of-living index" is supposed to tell us how much money income must change to enable consumers to maintain the *same standard of living.* But the CPI tells how much income must change to let people buy the *same goods.* Yet when prices go up, people systematically shift their purchases away from goods whose prices have risen the most, toward those whose prices have risen the least. This is how they protect themselves against inflation. By shifting toward chicken, whose price is stable, and away from steaks, whose price is skyrocketing, they need less income to maintain the same living standard than they would if they stuck doggedly to the same meat.

A plausible solution to the fixed-weight bias would be to shift to a moving-weight index. But besides being extraordinarily expensive for the BLS, if done right, this would not be very satisfactory. A moving-weight index has its own bias. It tells how much extra income people must get to protect themselves against increases in prices of the goods they actually consume, not those they would have consumed if prices had been stable. Thus, it doesn't take into account that they really prefer steaks to chicken. If COLA payments were based on a moving-weight index, they would not prevent a drop in the standard of living. No one, therefore, seriously argued that the BLS should switch to moving weights.

However, economists did successfully challenge the structure of weights used by the BLS, on the ground that it greatly overstated the relative importance of housing. This overstatement created problems because of the ex-

treme variability of mortgage interest rates.

The CPI is more a measure of the cost of *buying* than of the cost of *living*. For nondurable goods and services, the use of what is bought occurs at about the same time as their purchase. But buying and using durable goods are not the same. The use value of a durable is spread out over time—for housing, a very long time. Most economists would argue that the cost of using a house is the opportunity cost of what is invested in it, plus the out-of-pocket expenses of taxes, insurance, and upkeep. This is fairly well approximated by what the house would bring on the rental market, and the Commerce Department values the consumption of housing in just this way when it calculates GNP. The implicit price index for personal consumption in the GNP deflator in effect weights housing according to its *rental value*. But the BLS measures the cost of housing purchases minus resales, and weights housing costs in the CPI according to the average amount that people *spent* in the base year to buy houses, *plus* half of the *mortgage interest* they contracted to pay at the time of purchase. Most serious students of the CPI think that this systematically overstates the opportunity cost of using the housing stock. The commissioner of labor statistics, Janet Norwood, was persuaded by their arguments and decided to change how housing was handled in the CPI. The new treatment, to be adopted in 1983, will bring the BLS practice in line with that used by the Commerce Department. When the housing component of the cost of living is leading the inflationary pack, the new index will increase less rapidly than the old, to the injury of those who receive COLA payments, and to the benefit of those who pay them.

duced is evaluated in terms of its market prices in various years, and the resulting valuations are expressed as a ratio to their base-year value. The GNP deflator is only a little more complicated.

Real GNP and the GNP deflator

In the previous chapter, we discussed constant-dollar GNP. To understand more completely what "constant dollar" means, we must see how it is related to price increases.

The calculation of real or **constant-dollar GNP** uses several fixed-weight indexes. The Commerce Department first divides final demand into several components, each of which is a fine subdivision of consumption, investment, government purchases, exports, or imports. Then it calculates a price index for each of them, and uses this index to "deflate" the corresponding demand to get a constant-dollar quantity. *Deflation* means dividing a current-dollar GNP component by the same year's price index for that component. This puts each year's output in terms of the base year's prices. For instance, current-dollar consumption in 1980 was $1670 billion. The consumption deflator for the same year was 179.0 on a base of 1972 = 100. Consumption in 1980 measured in 1972 dollars was therefore $933 billion, since $933 = (1670 \div 179.0) \times 100$. Overall GNP in constant dollars is calculated by adding up the many deflated components.

One of the by-products of calculating real GNP is the "implicit GNP deflator." It

is called implicit because it is calculated from real GNP rather than directly. The ratio of current-dollar to constant-dollar GNP can be used to infer the general level of prices of the things that make it up. Specifically, the deflator is given by:

$$\text{Implicit GNP deflator} = \frac{\text{Current-dollar GNP}}{\text{Constant-dollar GNP}} \times 100.$$

Because the relative sizes of the components of real GNP vary from year to year, the weights implicit in the deflator also vary. The implicit GNP deflator is, therefore, mathematically equivalent to a *moving-weight price index*, unlike the CPI or PPI. Moving-weight indexes tend to show smaller rates of inflation than fixed-weight indexes. In effect, the market basket used to compare each given year's prices to the base year's prices is the market basket of that *given year*. Over time, consumers tend to reduce their purchases of goods that rise most rapidly in price, like gasoline. This means that a moving-weight index downplays the relative importance of the most rapidly inflating parts of the consumers' market basket, as compared to a fixed-weight index. The implicit deflator for the consumption component in GNP shows 4 percentage points less inflation between 1970 and 1980 than in a comparable fixed-weight index. Part of the criticism of the CPI in the box in this chapter hinges on the fact that it is a fixed-weight index.

You can get some idea of the relative sizes of the price and real output components of GNP by looking at Figure 6. Both current- and constant-dollar figures are plotted on the same axes. They coincide in 1972, which is the year for which the price index equals 100. The differing trends in the two series reflect the growth in prices over the period. Both these trends and the fluctuations that show up in Figure 6 illustrate the kinds of events—recessions, growth spurts, stagnation, inflation—that the following chapters will help you to understand.

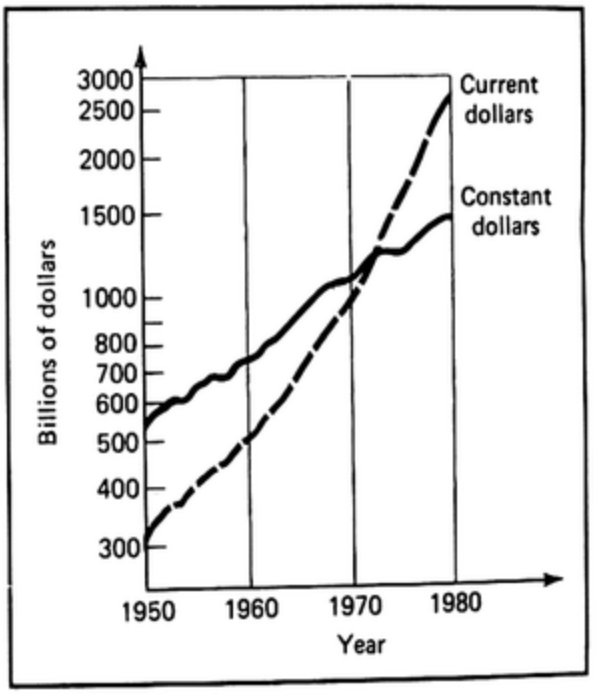


Figure 6 GNP in current and constant (1972) dollars

GNP measured in constant dollars shows a much less rapid growth trend than GNP measured in current dollars. From 1950 to 1980, current-dollar GNP increased by a factor of 9, constant-dollar GNP by a factor of 3. The difference was a threefold increase in prices. The two lines cross in 1972, which is the base year for the GNP deflator, the year in which it is equal to 100 and the current- and constant-dollar GNP figures are equal.

Source: *Economic Report of the President.*

Summary

This chapter has taught you to look at the economic system as an integrated whole, made up of household, business, government, and foreign sectors. These sectors are tied together by a network of flows of goods, services, expenditures, and incomes. The major things that you will want to remember are:

1. The volume of economic activity that takes place over some time period is usually measured by the gross national product (GNP), which is defined as the market value of the country's total output of goods and services in dollars per time period, minus the value of imports and of the intermediate goods used up in the production process.
2. There are several ways of measuring GNP, three of which are most useful in developing macroeconomic theory: by the *value added* in production; by *final demand* approach; and by *gross national income* (GNY).
3. The equivalence of these three approaches can be seen most clearly in an input-output table.
4. Government taxes and transfer payments bring about a major rearrangement of the income flows to and from the business sector.
5. Because GNP and GNY are necessarily equal, whenever one sector of the economy takes in income in excess of its final demand, and thereby runs a surplus, some other sector must run a deficit. The sum of sectoral surpluses (with deficits thought of as negative surpluses) must be zero.
6. When following GNP over time, it is necessary to disentangle price changes and quantity changes. This is accomplished by using price indexes. Price indexes chart the average price of a complex bundle of goods and services over time.

Key concepts

Factor services
Factor incomes
Household sector, business sector
Consumption goods
Investment goods
Intermediate goods
Foreign sector
Government sector
Product of government
Value added, final demand, gross national income (GNY)
Input-output table
Inventory investment
Gross fixed investment
Depreciation
Net fixed investment
Net national product
Net taxes
Purchases
Transfer payments
Gross business saving
Depreciation allowances
Net business saving
Personal income
Disposable income
CPI
PPI
GNP deflator
Base period
Constant-dollar GNP

Questions for review

1. Three of your friends are having a heated argument. One claims that GNP is measured by summing up final demand for goods and services. Another claims that GNP is the sum of value added by all firms. A third claims that GNP is actually the sum of wages, indirect taxes, and property income. Explain clearly and patiently why all three of them are correct.

2. Referring to Figure 4, the input-output study, consider the row and column representing the livestock industry. Explain why neither the figure for total output (40.704) nor that for total input (43.339) can be added directly into GNP.
3. Value added and final demand must be equal for the economy as a whole. Must they be equal for an individual industry? Why or why not?
4. Suppose that each sector of the economy shows a zero surplus. Then business firms decide to increase investment, creating a business sector deficit. Explain *how* this creation of deficit in the business sector will cause another sector (or other sectors) to show a surplus.
5. Why would a moving-weight price index tend to show a lower rate of inflation than a fixed-weight index?

24

Equilibrium of the Circular Flow

As you read and study this chapter, you will learn:

- what is meant by equilibrium of the circular flow
- how the incomes and demands of the sectors of the economy change as GNP changes
- why too low a level of GNP falls short of the demand for it, and too high a level exceeds the demand for it
- what relationships must be satisfied if demand and GNP are to be in equilibrium with each other

The next time you go to an amusement park, be sure to ride on the biggest roller coaster you can find, preferably one of the "corkscrew" type that has you upside down part of the time. If the roller coaster is well designed, you will hardly need a seat belt or a grab bar. The relationships among geometry, velocity, mass, gravity, and friction have been so carefully calculated by the designer that you will be continually pushed *into* your seat no matter how the track curves. The next time this happens to you, you might remember to admire the skill of the designing engineer. (If you find yourself being pushed *out of* your seat, change amusements parks.)

The theory of "mechanics" that enables engineers to design safe roller coasters was developed by Sir Isaac Newton about 300 years ago. He was fascinated by the motion of the planets and succeeded in formulating a set of "laws of motion" to explain them. He expressed these laws in a system of equations that describe the planetary orbits with great accuracy.

There is an obvious analogy between planetary motion and macroeconomics. Both planetary and economic systems are made up of interconnected parts. Their individual motions cannot be explained without referring to the relations among them. The major principle of macroeconomics is interdependence. It is thus not surprising that economics has taken over some of the spirit of Newtonian mechanics.

This chapter explains one of the most important theoretical constructs in macroeconomics—*equilibrium* in the circular flow. The planets stay in orbit because the forces tending to send them flying apart just balance those tending to make them crash together. A similar pattern is reproduced month after month when the forces in the circular flow of economic life just balance out. Like that of the planets, economic equilibrium is a kind of repetitive motion, not a state of rest.

You yourself may have been part of an equilibrium system while riding a roller coaster. When the car goes over a small hill at just the right velocity, its occupants have the momentary sensation of "weightlessness." They are neither forced into their seats nor thrown out of them. The car and the people are following consistent paths rather than fighting each other. They go over the top of the hill in formation. ***This relationship of compatibility is what economists have in mind when they think of equilibrium. Because the plans of the economy's major sectors are compatible, they produce, sell, buy, and consume goods and services "in formation."***

Consumption and equilibrium

When they try to understand a new principle, economists usually start by studying the *least* complicated problem that incorporates it. They then work step by step toward the world as it really is. As you learn to master the principles of equilibrium, think first about an economic system that has only two sectors—households and firms. This means ignoring government and foreign trade altogether. Depending on their disposable income, households decide what to spend on consumer goods. Business firms decide what to produce. Together, they set the levels of gross national product (GNP) and gross national income (GNY). But for the moment their investment demand is fixed, independent of GNP. And business saving is zero.

These are temporary ground rules, designed to make your first encounter with equilibrium theory successful. In the sections that follow, complications will be added, a few at a time. But for now, things are quite simple.

The theory of equilibrium has three main divisions. One, which we just discussed in the preceding chapter, covers how the interrelated sectors *fit together*—how the circular flow is laid out. The next deals with how the sectors *behave.* The final division combines the others and explains what the level of GNP must be for the behavior of the sectors to be consistent—for them to be *in equilibrium.*

The propensity to consume

The behavior of the household sector is a good place to begin the study of equilibrium, since it relates closely to your personal experience.

Do you remember when you got your first job? Your parents probably urged you to save some of your earnings, perhaps for some definite goal. But you must have felt that you had a right to spend most of your earnings on things you had always wanted, but couldn't afford when you were entirely dependent on your parents.

This propensity for people to divide their incomes between consumption and saving is characteristic of the household sector as a whole. Figure 1 shows data for

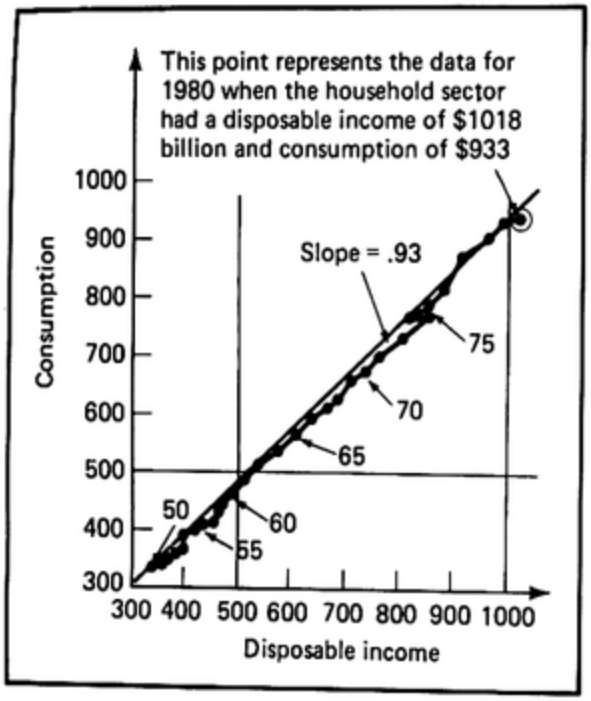


Figure 1 Consumption and disposable income 1950–1980 (in billions of 1972 dollars)

During the 30 years from 1950 to 1980, the disposable income of American households grew from about $350 billion to more than $1 trillion (in constant 1972 dollars). Yet the ratio of consumption to disposable income remained close to 93 percent throughout the period.

Source: *Economic Report of the President.*

the American economy over a 30-year period, comparing consumption and disposable income in constant dollars. Notice that the share of income consumed remained close to 93 percent over the whole period. The stability of this relationship is largely the result of averaging over millions of households. In any year, many, many households spend more than they take in. But other households save large amounts. From year to year, the consumption and savings pattern of any given household varies widely. This means that the identities of the big spenders and big savers change from year to year (although the very rich must surely save a lot nearly every year). But in the aggregate, much of the instability cancels out, leaving the overall saving ratio fairly stable over the long run.

Don't let this long-run stability keep you from noticing the short-run irregularities, however. They have a pattern too. If you look at Figure 1 very closely, you will see that whenever disposable income grows very rapidly (as from 1954 to 1957 and from 1965 to 1968), the share of income consumed tends to drop from one year to the next. When disposable income hardly grows at all (as from 1953 to 1954 and from 1973 to 1975), both consumption and the share of income consumed rises faster than income itself.

This relationship stands out more clearly in Figure 2, which graphs year-to-year *changes* in both consumption and disposable income. Judging from the line that best fits the data, consumption rises by less than income when income goes up by a lot, and by more than income when income rises by very little. When the line is projected back to a zero change in income, the consumption intercept is positive. This

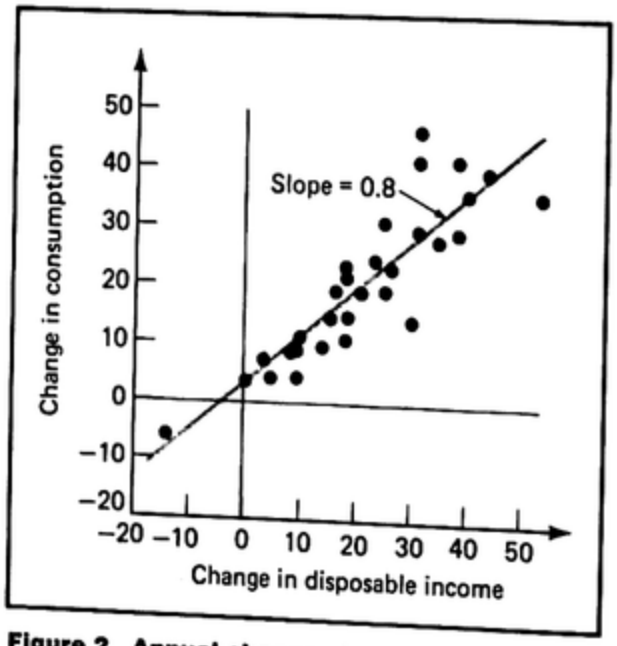


Figure 2 Annual changes in consumption and disposable income 1950–1980 (in billions of 1972 dollars)

Year-to-year changes in consumption are fairly closely related to changes in disposable income. As the diagram shows, consumption tends to rise even when income does not. Each $1 billion rise in disposable income also adds about $0.8 billion to the rise in consumption.

Source: *Economic Report of the President.*

implies some tendency for consumption to rise even when income is stationary (remember, the axes measure changes). During extended periods of rapid growth in income, consumption lags behind. When income stops growing, consumption catches up.

If you think about yourself, you may be able to guess some of the reasons that economists have advanced to explain this short-run pattern, in which consumption lags and then catches up. The most plausible is that consumption is regulated partly by habit and partly by calculation. When people's income rises rapidly, their consumption lags, simply because it takes some time for their consumption habits to react to the reality of their higher income. When income growth slows, habits have time to adjust. For people whose income has dropped, the problem of changing habits is painful, as you may know first hand, so that the lag on the down side may be quite long. Meanwhile, saving absorbs the brunt of the drop in income. If you have ever had to suffer through a big drop in family income, you know how hard the adjustment can be. You keep hoping things will get better, so that you won't have to sell the house or cut back to just one car.

The relationship between consumption *(C)* and disposable income *(YD)* is usually called the ***consumption function.*** It can be written in the form of a table, an equation, or a graph. Since personal saving *(SP)* has to equal disposable income minus consumption, knowing the consumption function also tells you the relationship between saving and disposable income. This means that every consumption function is paired with a ***saving function.*** As with the consumption function, the saving function can be written as a table, an equation, or a graph. Figure 3 illustrates consistent consumption and saving functions in all three ways. You should study it long enough to see both the equivalence among the three ways of expressing these relationships and the consistency between the consumption and saving functions.

A lot of discussion is greatly simplified by using a specialized vocabulary to describe the consumption and saving functions. Two sets of terms are helpful. First, there are the ***average propensity to consume*** and the ***average propensity to save.*** These averages are the ratios of consumption and saving to disposable income:

$$\text{Average propensity to consume} = \frac{\text{Consumption}}{\text{Disposable income}}$$

$$\text{Average propensity to save} = \frac{\text{Saving}}{\text{Disposable income.}}$$

They are often called the APC and the APS. According to the consumption and saving functions given in Figure 3, the APC and APS, *evaluated at a disposable income of 1,000,* are given by:

$$\begin{aligned} APC &= 920/1000 \\ &= .92 \\ APS &= 80/1000 \\ &= .08. \end{aligned}$$

At a disposable income of 600, they are given by:

$$\begin{aligned} APC &= 600/600 \\ &= 1 \\ APS &= 0/600 \\ &= 0. \end{aligned}$$

For these particular consumption and saving functions, the APC and APS vary as disposable income varies. Since consumption rises and falls in a smaller proportion than income, the APC falls as income increases and rises as income falls. Necessarily, then, the APS rises as income rises and falls as income falls. Because consumption and saving add up to income, the APC and the APS add up to 1.

The APC and APS have two companion concepts, the ***marginal propensity to consume (MPC)*** and the ***marginal propensity***

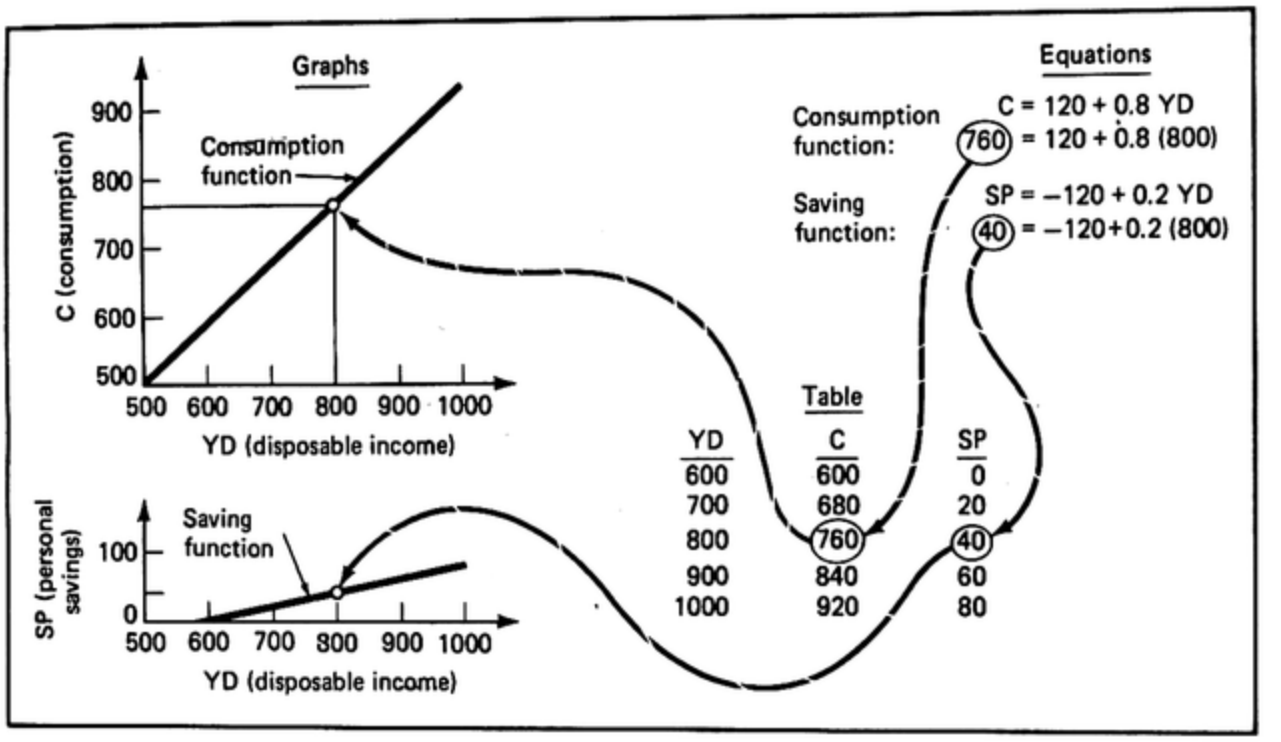


Figure 3 Consumption and saving functions

These hypothetical equations, graphs, and the table all illustrate the same consumption and saving functions. Notice that according to all three, a disposable income of 800 is divided into 760 of consumption and 40 of saving.

to save (MPS). These are also ratios, but they involve changes rather than levels. They are defined by:

$$MPC = \frac{\text{Change in consumption}}{\text{Change in income}}$$

$$MPS = \frac{\text{Change in saving}}{\text{Change in income}}.$$

Refer back to Figure 3 and look at the changes in consumption and saving between different values of disposable income. For example, calculate the MPC and MPS between income levels of 600 and 1,000:

$$MPC = \frac{920 - 600}{1000 - 600} = \frac{320}{400} = .8$$

$$MPS = \frac{80 - 0}{1000 - 600} = \frac{80}{400} = .2.$$

You can verify from looking at other pairs of income levels that the MPC and MPS in Figure 3 are 0.8 and 0.2 throughout. This follows from representing the consumption and saving functions by straight lines with constant slopes. It would not be true if the consumption and saving functions had been represented by curves of changing slope. ***It must always be true, however, that the MPC and MPS add up to 1, no matter what the level of income. The change in consumption and the change in saving must add up to the change in income.***

In real life, the APC and APS are used for one set of problems, the MPC and MPS for another. If you are thinking about growth and the accumulation of capital, the division of income between consumption and saving is critical. The APC and APS are the right numbers to look at. But if you are interested in the short-run response of demand to changes in income, the MPC and MPS are what you want to

know because they measure the reaction to change.

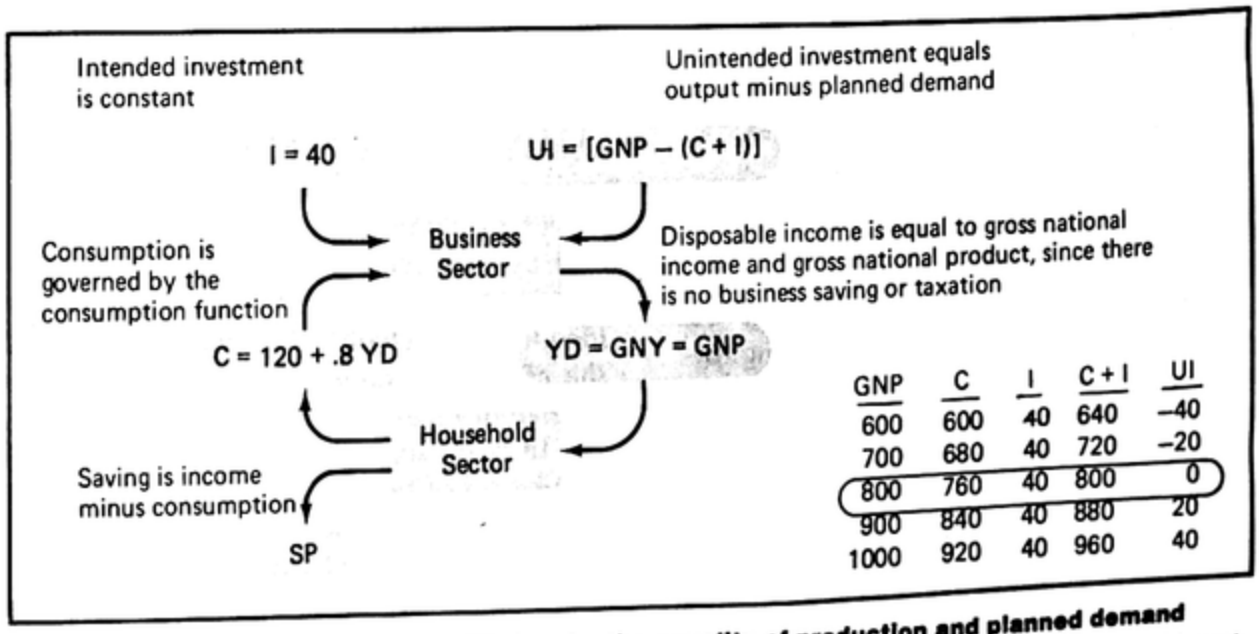
Combining the two sectors

If you have taken a beginning course in physics or in calculus for students of physical science, you have probably studied the parabolic arc of a rocket in free flight, what an American novelist once called "Gravity's Rainbow." This is a good example for starting physics because the gravitational pull of the rocket is so small relative to that of the earth that it can be ignored without introducing measurable error into calculations of the arc.

In the present example the business sector is like the earth. It generates a pull on the household sector. By setting GNP, it sets GNY. Since the business sector in this simplified example does not save, GNY and YD (disposable income) are equal. In effect, therefore, the business sector determines household income and (indirectly) consumption. It also determines its own rate of investment, so that it directly and indirectly controls final demand. But for the moment, its production decisions are assumed to be independent of final demand. The earth is not pulled around by the rocket.

Figure 4 shows you how to determine equilibrium GNP. The consumption function built into this example is already familiar to you. It is the same one that is illustrated in Figure 3. Since business saving and taxes are assumed to be zero, disposable income is identical to gross national income (GNY), which equals GNP. The first two columns of Figure 4's table are therefore identical with those of Figure 3's.

Figure 4 assumes a level of investment *(I)* of 40, independent of the level of GNP. In the last chapter, the symbol *I* was used to denote actual investment. In studying equilibrium, it is important to distinguish between planned and unplanned invest-



GNP	C	I	C + I	UI
600	600	40	640	–40
700	680	40	720	–20
800	760	40	800	0
900	840	40	880	20
1000	920	40	960	40

Figure 4 The determination of equilibrium by the equality of production and planned demand
The circular flow is in equilibrium when the consumption of the household sector *(C)* plus the planned investment of the business sector *(I)* is just equal to GNP. In this example, a GNP of 800 will generate 800 of demand—760 of consumption and 40 of planned investment. At any other level of GNP, demand will be different from GNP, and firms will find themselves investing more (if GNP is too big) or less (if GNP is too small) than they planned to invest. The column labeled *UI* measures the amount of unplanned investment in inventories of unsold goods. It equals output minus planned demand.

ment. You should now think of I as the level of ***planned investment***—what firms will actually invest if they sell exactly what they produce. But if they cannot do this, they will invest more or less than they planned to. This ***unplanned investment*** *(UI)* will take the form of unintended gains or losses of inventory. It can be positive or negative. If GNP equals $(C+I)$, then firms will invest what they planned to invest. But if GNP is bigger than planned demand, the unsold goods will pile up in inventories, and unplanned investment will be positive. If GNP is smaller than planned demand, this unplanned investment in inventory will be negative. Firms will lose inventories they wish they had.

The arithmetic of Figure 4 is straightforward. Suppose that firms produce a level of GNP of 600. Consumption will be given by:

$$C = 120 + .8\,(YD)$$

$$= 120 + .8\,(600) = 600.$$

Since planned investment is 40, total planned demand is 640. Notice that this planned demand is bigger than GNP. Business will lose 40 in inventories of consumer goods that they would prefer to have in stock. Actual investment will be zero because the unplanned inventory drop of 40 will offset the 40 of planned investment. Clearly, then, 600 cannot be an equilibrium level of GNP.

Now look at what happens at a GNP of 1,000. Consumption is 920 and planned investment is 40, so that total planned demand is 960. That will not buy up a GNP of 1,000. Inventories will absorb the leftover 40 as unplanned investment. Again, the business sector will not meet its goals.

Finally, look at what happens if GNP is 800. Consumption is 760. Since planned investment is 40, total planned demand is 800. Because this just matches production, there is no unplanned investment in inventory. This is therefore the equilibrium GNP, the level at which the plans and reactions of the sectors are consistent with one another. Mama Bear's bed fits Goldilocks just right.

Before going on to the next section, it would help to work out the numbers in the table of Figure 4, at least far enough so that you are sure you understand all of them. Start with GNP and work your way clockwise around the circular flow, calculating C, $C+I$, and finally $UI = GNP - C+I$.

Saving, investment, and the sectoral surpluses

There are two other ways of expressing the conditions of equilibrium. One way focuses on saving and investment, or more generally, withdrawals from and additions to the circular flow. The other focuses on the surpluses of the various sectors whose transactions make up the circular flow. Studying these other ways is not just a "make-work project." Many students find it helpful to look at equilibrium from alternative vantage points.

To see what these other conditions are, and why they are equivalent to the first, start with the equality between planned demand and GNP:

$$GNP = C + I.$$

Now notice that in general, GNP equals GNY, and that in the simplified example, $GNY = YD$. So a restatement of the equilibrium condition is:

$$YD = C + I.$$

Then recall that YD is divided between C and SP, so that

$$C + SP = C + I$$

is another way of writing the same condition. Finally, cancel C from both sides, so

that what is left as an equilibrium condition is:

$$SP = I.$$

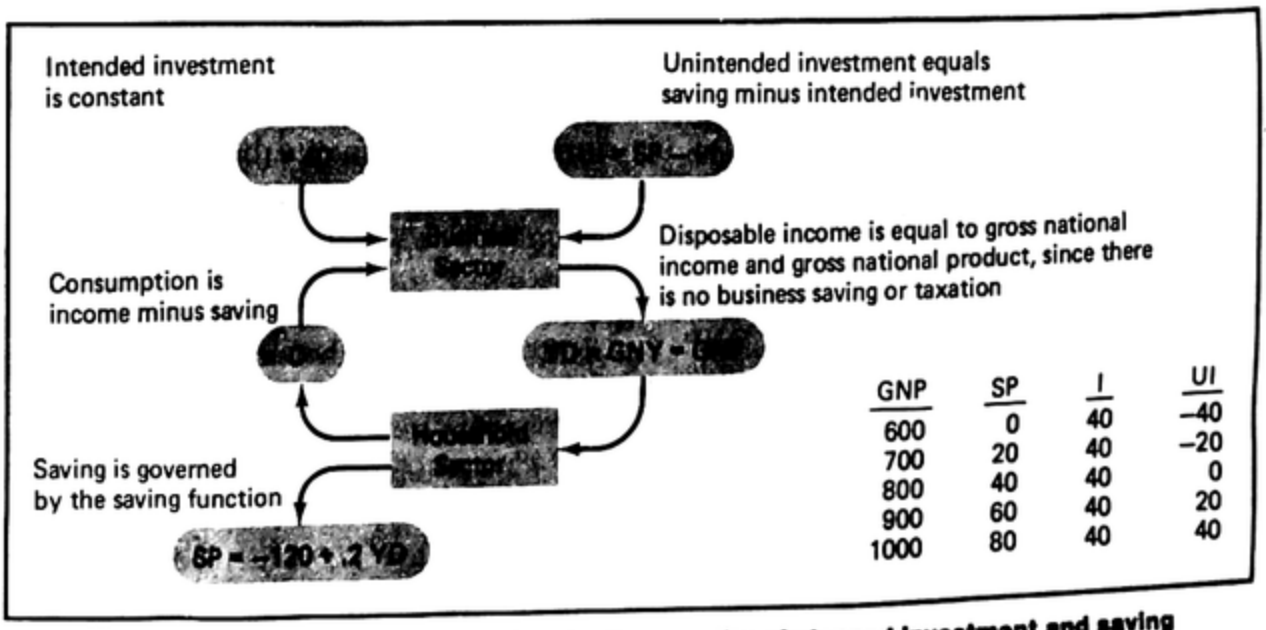
Apparently, equality between saving and planned investment is equivalent to equality between production (GNP) and planned demand *(C + I)*.

If you think for a moment, you will see that this must be true. Households get a certain disposable income, which equals GNP. They save part of this *(SP)*, which is not returned to the circular flow as demand for goods and services. It is a *withdrawal* from that flow. If GNP is to be purchased, this withdrawal must be offset by an *addition* to it. When $SP = I$, the addition coming from planned investment exactly offsets the withdrawal as saving, and planned demand as a whole *(C + I)* is exactly matched with GNP.

The relationship that focuses on the sectoral surpluses is trivial in the simple two-sector example with no business saving, since if $SP = I$, the *planned sectoral surpluses* obviously add up to zero and are consistent with one another. The planned household surplus is *SP*. If business does not save, then *I* is its planned deficit. Equality between *SP* and *I* means that the planned surplus of one sector is completely offset by the planned deficit of the other. Remember that *actual* surpluses and deficits must always cancel out. But *planned* surpluses and deficits will only cancel out when the plans are consistent, that is, when the circular flow is in equilibrium.

To see why inequality between *SP* and *I* implies disequilibrium, turn to the circular flow diagram in Figure 5. The only difference between Figure 5 and Figure 4 is one of emphasis. Households save out of disposable income an amount that is given by their saving function. Since the saving function in Figure 5 is the companion of the consumption function in Figure 4, the same behavior is expressed in both figures.

You should think about Figure 5 in this way: Households receive GNY, save some portion of their receipts, and return



GNP	SP	I	UI
600	0	40	−40
700	20	40	−20
800	40	40	0
900	60	40	20
1000	80	40	40

Figure 5 The determination of equilibrium by the equality of planned investment and saving
In this way of looking at equilibrium, the focus is on personal saving and investment. If savers withdraw more savings from the circular flow than firms are investing, then the flow of planned demand is insufficient to buy up GNP. Firms are forced to accumulate unwanted inventories. The flow is in equilibrium when saving exactly offsets planned investment.

the rest to the circular flow in the form of consumption. The circular flow will be in equilibrium if the saving households withdraw from the circular flow is exactly offset by the planned investment that business adds to it. But if planned investment differs from saving, then there will be unplanned investment, either positive or negative. As you can see from Figure 5, saving and planned investment are equal at a GNP of 800. This is exactly the level of GNP at which planned demand and GNP are equal. If you carefully compare the numbers in Figures 4 and 5, you will see that because they are two different ways of looking at the same circular flow, they must give the same equilibrium.

Income and spending of the business sector

As you just learned, one way of locating equilibrium is to find the level of GNP at which the surplus of the household sector *(SP)* equals the planned deficit of the business sector *(I)*. In Figure 5, this was particularly easy because the business deficit was fixed. Business saving was assumed to be zero, and planned investment was fixed at 40. You had only to find the level of GNP at which household saving was 40.

In fact, the world is not quite so simple. When businesses save, their deficit is the *difference* between investment and business saving. Both planned investment and gross business saving change systematically in response to changes in GNP. This makes the arithmetic of equilibrium more complicated than Figure 5 suggests. But the principle remains simple. ***The circular flow is in equilibrium when the business sector's net addition to the flow of spending (its planned deficit) just offsets the amount withdrawn as household saving (a surplus). At this point, because of the offsetting surplus and deficit, planned demand is exactly what is needed to buy up GNP.***

Variations in business saving

While the income of the household sector comes largely from the sale of inputs such as labor and capital services, the income of the business sector comes from the sale of output. Since purchases and sales of intermediate goods cancel out for the business sector as a whole, all this income can be traced directly or indirectly to the sale of final products. Part of business income is spent or paid out as wages, rents, interest, dividends, taxes, and purchases of intermediate goods. Some portion of a firm's income, however, is not paid out to households or to other firms, but is retained by the firm. This portion is called ***gross business saving (SB).*** Since some of the income generated by production never reaches the household sector, disposable income would be smaller than GNY, even if there were no taxes.

What does a firm do with its gross business saving, which might be thought of as its disposable income? A large part of it is used to finance investment in fixed capital and inventories, which adds to final demand. But since the business sector usually runs a deficit, its investment is larger than its saving. The difference must be covered by outside financing. Financially secure firms can raise outside funds with ease. Shaky enterprises largely depend on internal funds—gross business saving.

Gross business saving is the sum of two components: retained earnings and depreciation allowances. ***Retained earnings*** are profits that are not paid out in dividends. Some funds must also be retained by firms to replace machinery and equipment that has worn out or depreciated with use. These additional funds are called ***depreciation allowances.***

There is a fairly close relationship both between business saving and gross national income. Depreciation allowances increase slowly over time, since they are linked to the accumulated total of past investment in durable capital. Retained earnings, on the other hand, fluctuate widely with the business cycle because they are directly linked to profits. Since the relative stability of depreciation allowances and the instability of retained earnings (net business saving) largely balance out, gross business saving absorbs about 12 percent of GNY in both good times and bad. It fluctuates erratically in the neighborhood of this figure, but the percentage of GNY going into business saving does not systematically rise and fall with GNY. To borrow the terminology of the household saving function, the marginal and average propensities to save are about equal.

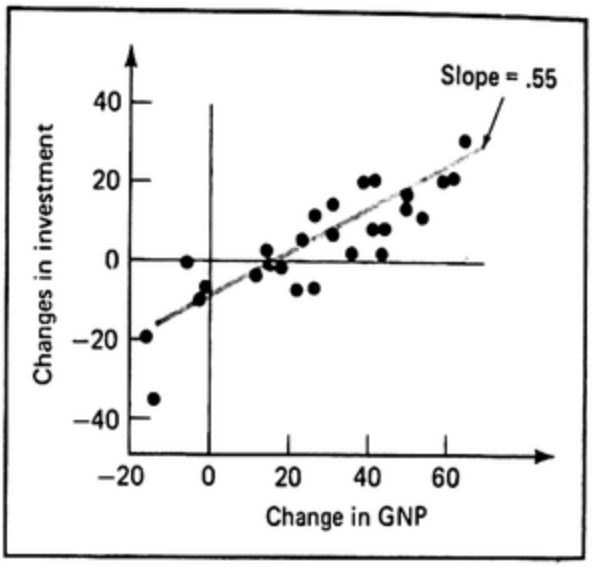


Figure 6 Annual changes in constant-dollar investment and GNP 1950–1980 (in billions of 1972 dollars)

Recent data show a strong association between changes in GNP and changes in investment. As the text points out, however, it is very difficult to measure cause and effect, since each of these variables affects the other.

Source: *Economic Report of the President.*

Variations in planned investment

The expenditure component of the business deficit is investment—one of the major sources of final demand. This is a very volatile component of GNP, especially investment in inventories. As Figure 6 shows, changes in investment are nearly as big in some years as changes in total GNP. The figure displays quite a strong association between changes in I and GNP, with a slope ($\Delta I/\Delta GNP$) equal to .55, based on constant-dollar quantities.

Actually, the relation between I and GNP is not as clear-cut as it looks in Figure 6. Seeing why will show you how easy it is to be misled by data coming from an *interdependent* system like our economy. First, remember that investment is one of the components of final demand and GNP—in fact, the most widely fluctuating component from year to year. This means that I and GNP *must* move together simply because one is part of the other. It is possible to make allowance for this by concentrating on how much investment varies with *other* parts of final demand, and this does reduce considerably the degree to which investment seems to respond to GNP changes. But even after making this allowance, there is still a mixup of cause and effect. This happens because an increase in investment raises production in the investment goods industry. Recall the input-output process: As more is produced and more inputs are bought or hired, income throughout the economy increases. Some of this increased income is consumed, calling for a further increase in production and income. This process of action and reaction is called the *multiplier* and will be discussed in the next chapter. For now, simply note that changes in investment can cause changes in consumption. This interaction makes it very hard to determine how much of a change in I is the *effect* of changes in GNP and how much is the *cause*.

Don't be upset by the fact that investment affects GNP and GNP affects investment. Reciprocal cause and effect are

really very common. When you arm wrestle, your arm determines the position of your opponent's, and his or hers determines the position of yours. It is a case of simultaneous equations.

In fact, there are several reasons to suppose that much of the fluctuation in I occurs for reasons other than immediate GNP changes:

1. Inventory investment has a complicated pattern. It is partly planned and partly unplanned. A rise in demand makes firms want to have more inventories. However, if the demand increase is unexpected, inventories are first drawn down to supply the unexpected demand. To bring inventories back up to the desired level, there is now a burst of accumulation. Once inventories have been restored to the desired level, investment will decline even without a decline in demand.
2. Fixed investment tends to require long planning periods and to be made on the basis of long-term profitability calculations. Although a current rise in GNP may make the managers of firms more optimistic, its immediate impact on long-term investment will be small. Determining long-term expectations is a fairly complex matter that depends on much more than current GNP.
3. Some long-term investments, especially construction, are heavily influenced by the cost and availability of credit. Thus, to understand investment changes, you have to look at money and credit, not just GNP.

For these and other reasons, investment changes cannot be linked simply to changes in GNP, or to any other single influence. Most of the rest of the book will treat investment as though it were largely *autonomous* or *independent* of income. This does not mean that its fluctuations are completely random, or that economists do not understand what causes changes in investment. It simply means that investment is not *dependent* on GNP in the same direct, simple way that consumption is.

To the extent that investment does depend directly on GNP, however, this dependence influences how equilibrium is determined. In the next few pages, some dependence of investment on GNP will be included in the analysis of equilibrium. This will help you understand the qualitative importance of this influence, even if its quantitative importance cannot be measured with precision.

Equilibrium and the business deficit

If business investment and saving vary as GNP varies, so does their difference (the business deficit), unless by coincidence its two components offset each other. To illustrate how this variation influences the equilibrium of the circular flow, we have to make some specific assumptions about how investment and business saving vary. Suppose that they obey the following relationships:

$$I = -16 + 0.15\,GNP$$
$$SB = 0.10\,GNY.$$

According to these relationships, at $GNP = GNY = 320$, business saving exactly balances investment at a level of 32. At any higher level of GNP, the business sector runs a deficit, since every increase in GNP raises investment by more than it raises business saving. Don't attach any particular significance to these numbers in themselves. They have been chosen to make the arithmetic of equilibrium work out easily. What is important is that you understand the logic of the equilibrium calculations.

These calculations are shown in Table 1, which is much like the table included in Figure 5. The calculations necessary to work out the entries in this table are based

Table 1 ***Equilibrium with a changing business deficit***

GNP	I	SB	I-SB	YD	SP	UI
600	74	60	14	540	−12	−26
700	89	70	19	630	6	−13
800	104	80	24	720	24	0
900	119	90	29	810	42	13
1000	134	100	34	900	60	26

According to the assumptions built into this table, 800 is the equilibrium level of GNP. At that level, the business deficit matches the household surplus.

To verify this, calculate all of the entries in the table at $GNP = 800$:

$$I = -16 + 0.15\,GNP = -16 + 120 = 104$$
$$SB = 0.10\,GNY = 80$$
$$I\text{-}SB = 104 - 80 = 24$$
$$YD = GNP - SB = 800 - 80 = 720$$
$$SP = -120 + .20\,YD = -120 + 144 = 24$$
$$UI = SP - (I\text{-}SB) = 0.$$

To be sure you understand this, verify the entries in at least one other line of the table.

on the investment and business saving functions in the previous paragraph, and on the personal saving function used in Figure 5. Notice, however, that when gross business saving is positive, as it is in Table 1, disposable income is smaller than GNY by the amount of the business saving that never gets into the hands of households. This is why every entry for personal saving in Table 1 is smaller than the corresponding entry in Figure 5. (Can you figure out why the difference in each case is 2 percent of GNP? Hint: 2 percent is 20 percent of 10 percent.)

You can see that a variable business deficit causes no new conceptual problems. The household surplus *(SP)* is smaller than the business deficit *(I-SB)* at a low level of GNP. As GNP gets higher, the household surplus rises relative to the business deficit. At a GNP of 800, they are equal. Unintended investment is zero, and the circular flow is in equilibrium.

Before going on to the next section, you might want to reassure yourself that the equality of the household surplus and business deficit in Table 1 is equivalent (1) to an equality between saving and investment, and (2) to an equality between planned demand and GNP. Indeed it is. At a GNP of 800, total saving is $SB + SP = 80 + 24 = 104$. This, of course, equals planned investment. To calculate planned demand, you need only subtract personal saving from disposable income to get consumption, and then add planned investment to get planned demand. At a GNP of 800, consumption is $720 - 24 = 696$. Adding planned investment of 104 gives a total planned demand of 800, just equal to GNP, as advertised.

Government, foreign trade, and equilibrium

This final section of the chapter fits the government and foreign trade sectors into the equilibrium of the circular flow. Now that you have worked through the previous two sections, you could almost write this one yourself—if you had the facts. You would only need to mesh the government and foreign sector surpluses and deficits with those of the two private domestic sectors. ***The condition of equilibrium is simply that the planned deficits and surpluses of the four sectors must balance one another out. If they do not, there is unplanned investment or disinvestment. If they do, then planned demand equals GNP.***

The government

Remember how the government fits into the circular flow. Its expenditures fall into two quite different categories. The first is ***purchases of goods and services (G),*** either from private firms or directly from public employees. These purchases are part of final demand and measure the product of government. The second kind of expendi-

tures are *transfer payments* (Social Security, unemployment, welfare, and similar payments), which supplement incomes paid by the private sector. These transfers increase household income, but do not directly purchase goods and services. Government receipts are mainly *taxes* which reduce both profits and disposable income.

In analyzing the circular flow, it is simpler to group taxes and transfers together into a single figure, called ***net taxes (TN),*** defined as taxes minus transfer payments. This is the *net* income drain from the circular flow to the treasuries of federal, state, and local governments. This simplification lets GNY be expressed as the sum of three parts—disposable income, business saving, and net taxes. Each of these three numbers is conveniently tied to one of the domestic sectors of the economy.

Net taxes are large and very responsive to changes in GNY. On average, net taxes at all levels of government absorb about 20 percent of GNY. But when GNY changes, net taxes take about 25 percent of the change. Thus, the *share* of GNY going to net taxes rises when income rises and falls when it falls. One main reason for the high income sensitivity of net taxes is that transfers move in the opposite direction from GNY. If, for example, GNP and GNY *fall,* unemployment compensation *rises,* and therefore taxes minus transfers (including unemployment compensation) *fall a lot.* Another is that short-run variations in GNY fall disproportionately hard on corporate profits, which are taxed at nearly a 50 percent rate. Because of their high cyclical variability, net taxes act automatically to cushion disposable income from the impact of changes in GNP. Thus, many of the provisions of the laws that govern taxes and transfers are called "built-in stabilizers." This stabilizing property of net taxes will be a main focus of the chapters on stabilization policy.

Unlike tax revenues, government purchases of goods and services are very largely autonomous. They do not vary strongly with income. State and local purchases are closely tied to population trends, since they mainly go to finance education, police, fire, sanitation, and other public services. These expenditures are not very cyclically variable. Federal purchases are more erratic, but most of the volatility comes from military expenditures. The fluctuations in defense expenditures have in modern time been a major *source* of disturbance in the civilian economy, but they are the cause of instability, not the consequence. For these reasons, it is best to treat government purchases as autonomous. This is not quite true for local government. For example, when the Chrysler Corporation closed its "Dodge Main" plant for good in 1979, idling thousands of workers, the city of Hamtramck, Michigan, had to lay off 80 of its own employees because of a loss of tax revenues. In general, though, treating government purchases as autonomous is more often correct than linking them to income.

The budget surplus or deficit of the government sector as a whole is the difference between purchases and net taxes. Purchases are largely autonomous. Net taxes rise sharply as income rises. Therefore, the budget of the government sector swings toward surplus (taxes bigger than expenditures) when GNP and GNY go up, and toward deficit (expenditures bigger than taxes) when GNP and GNY go down. Government policymakers directly control purchases, but they can only write tax and transfer laws. When GNP changes, so does the surplus or deficit, even though there have been no changes in the laws.

Foreign trade

Foreign trade, like the government budget, enters the circular flow as a component

(exports) that is largely independent of GNP, and as another component (imports) that varies directly with GNP. Our *exports (X)* are bought by foreigners. What they buy is influenced by their GNP and by relative prices, but not by our GNP. Our imports *(M)*, however, are largely inputs into our own production process. As such, they go up when our GNP increases, just as the use of domestic inputs goes up. Since *imports* are subtracted from exports to get *net exports*, net exports fall as the GNP and imports rise. This means that expansion in domestic production results in a trade deficit or a smaller surplus. At the margin, imports rise by about 15 percent of any increase in GNP. This is nearly twice what they are on the average, so that downswings in GNP usually bring a big "improvement" in the balance of U.S. trade with the rest of the world.

In analyzing surpluses and deficits in the circular flow, you should think of a U.S. trade deficit (imports bigger than exports) as a surplus for the foreign sector. If our imports are bigger than our exports, the dollar income of the rest of the world is bigger than its dollar expenditures. This is a surplus for the rest of the world. Since our imports rise sharply as our GNP goes up, the dollar income of the rest of the world rises relative to its expenditures, and the trade balance of the rest of the world swings toward surplus. When our GNP goes down, the trade balance of the rest of the world swings toward deficit.

Equilibrium of all four sectors

Fitting the household, business, government, and foreign sectors together into an equilibrium pattern may seem a lot like assembling a Chinese puzzle. You saw the pieces come apart with your own eyes, but there seems to be no way to reassemble them.

The pieces do fit together. But before trying to reassemble them, it will help you to look at them all at once. Table 2 reviews the components of the various sectoral surpluses and reminds you how they move. You should go over it to make certain you remember how things work.

Table 2 ***Behavior of sectoral incomes and demands***

Sector	Receipts	Planned demand	Surplus/deficit
Household	*YD*—disposable income (procyclical)	*C*—consumption (procyclical)	*SP*—personal saving (procyclical)
Business	*SB*—gross business saving (procyclical)	*I*—planned investment (procyclical)	*(SB-I)*—business surplus (?)
Government	*TN*—net taxes (procyclical)	*G*—government purchases (autonomous)	*(TN-G)*—government surplus (procyclical)
Foreign	*M*—imports (procyclical)	*X*—exports (autonomous)	*(M-X)*—U.S. trade deficit or foreign trade surplus (procyclical)

The receipts of all four sectors are *procyclical*—they move in the *same* direction as GNP. The planned demands are either procyclical or autonomous. The surpluses of the household, government, and foreign sectors are procyclical, but that of the business sector may go either way, depending on whether investment or business saving dominates.

Table 3 ***Equilibrium and disequilibrium in the four-sector economy: the three equivalent conditions***

GNP	SB	TN	YD	SP	C	I	G	X	M	UI
900	110	175	615	55	560	145	220	80	55	-50
1000	120	200	680	70	610	160	220	80	70	0
1100	130	225	745	85	660	175	220	80	85	+50

At a GNP of 1,000, each of the three equivalent conditions of equilibrium is satisfied. To see this, start from the first condition.

1. *GNP equals planned demand:*

$$GNP = C + I + G + (X - M)$$
$$1000 = 610 + 160 + 220 + (80 - 70).$$

Remember that $GNP = GNY$, so substitute $GNY = YD + SB + TN$ for GNP, and $C + SP = YD$, cancel the C's, and move M to the left:

$$\overbrace{GNP = GNY}$$
$$\overbrace{YD + SB + TN} = C + I + G + (X - M)$$
$$\overbrace{C + SP} + SB + TN + M = C + I + G + X$$

This gives

2. *The withdrawals from the circular flow balance the additions:*

$$SP + SB + TN + M = I + G + X$$
$$70 + 120 + 200 + 70 = 160 + 220 + 80$$
$$460 = 460.$$

Now simply rearrange this to get

3. *The sectoral surpluses add up to zero:*

$$SP + (SB - I) + (TN - G) + (M - X) = 0$$
$$70 + (120 - 160) + (200 - 220) + (70 - 80) = 0$$
$$70 - 40 - 20 - 10 = 0.$$

Table 3 illustrates how the pieces fit together. It has been set up to make the arithmetic easy, but the relative sizes of the numbers are fairly realistic. For example, the division of GNY into *SB*, *TN*, and *YD* is roughly 12 percent, 20 percent, and 68 percent. This is about right for the U.S. economy. Consumption is about 90 percent of disposable income, but the *MPC* is about 0.8. Again, about right. For every 100 change in GNP and GNY, the following are the assumed changes in sectoral receipts and demands:

$$\left.\begin{aligned} SB &= 10 \\ TN &= 25 \\ YD &= 65 \end{aligned}\right\} \text{ Changes in } GNY = 100$$

$$\left.\begin{aligned} SP &= 15 \\ C &= 50 \end{aligned}\right\} \text{ Changes in } YD = 65$$

$$I = 15$$
$$M = 15.$$

Both G and X are assumed to be autonomous at roughly realistic levels relative to GNP.

Table 3 should be looked at in much the same way as Table 1. You should check that it incorporates the marginal propensities just given. You should also see that the distribution of GNY into *SB*, *TN*, and *YD*, and of *YD* into *C* and *SP*, match up with the average figures given earlier in the text. This will help to fix the right general picture of the economy in your mind.

Note also that like Tables 1 and 2, Table 3 shows a single equilibrium level of GNP and two disequilibrium levels. Since a GNP level of 900 intended demand is bigger than production, someone's plans cannot be realized at that level. Similarly, at 1,100, the entire production will not willingly be bought. If this level of output is produced, firms will accumulate 50 in un-

wanted inventories, and I will not equal its planned level. But at a GNP level of 1,000, everything works out. You may verify this by looking at the equalities at the bottom of the table, which show that the condition of equilibrium is satisfied in each of the various ways it can be written.

Summary

Equilibrium theory is not the most interesting part of macroeconomics. Picture a dozen 6-year-olds sitting in 12 chairs. They fit, exactly one per chair. The fun doesn't start, however, until the music begins, everyone gets up, and someone takes one of the chairs away. That is the subject of the next chapter—what happens when equilibrium is upset by a change in the behavior of one of the sectors. Equilibrium analysis is mainly a prerequisite for the theory of disturbance and adjustment, of which it is one of the building blocks.

Before you start on the next chapter, be sure you understand the following major points that have been developed in the present chapter:

1. The circular flow of production, income, and demand is in equilibrium when the plans of the participants are consistent with one another.
2. One way of stating this consistency is that planned demand of the four sectors combined equals GNP. If it does not, then inventories will pile up or run down despite what producers want.
3. Another way of stating the condition for equilibrium is that the withdrawals from the circular flow for saving, taxes, and imports must be offset by additions from planned investment, government purchases, and exports.
4. Still a third way is that the planned surpluses of the four sectors must add up to zero.
5. In any given year, the four sectors combined will plan a deficit at a relatively low level of GNP and a surplus at a relatively high level. If GNP is too low, demand will be larger than GNP, and there will be unintended disinvestment in inventories. If GNP is too high, demand will be smaller than GNP, and unwanted inventories will accumulate. Somewhere in between lies the equilibrium level of GNP, at which planned demand just equals GNP.

Key concepts

Consumption function
Saving function
Average propensities to consume and save (APC, APS)
Marginal propensities to consume and save (MPC, MPS)
Planned and unplanned investment
Autonomous demand
Retained earnings
Gross business savings

Questions for review

1. Studies show that consumption varies directly with disposable income. Yet, when income rises or falls, changes in consumption seem to lag behind and then catch up. Explain this pattern of consumption changes adjusting slowly to income changes.
2. A friend of yours is puzzled. She knows from reading the chapter that the average propensity to consume (or APC)

falls as income rises. Yet, she also knows that richer people spend more than poorer people. She feels that these pieces of information are incompatible. Explain her confusion and show why a falling APC and greater spending by richer people are not incompatible.

3. If GNP influences investment positively, could investment fall in a year when GNP was increasing? Explain.
4. Suppose that you are reviewing this chapter with a friend, and make the statement that planned demand equal to GNP is one equilibrium condition. Your friend replies that if that is the case, the economy must always be in equilibrium. His reasoning goes like this: Demand must always equal GNP. This is so because unsold goods are added to firms' inventories and are counted as part of demand. Thus, since demand must always equal GNP, the economy must always be in equilibrium. Right? Explain your answer.
5. Why is unplanned inventory accumulation or decumulation such an important clue as to whether the existing level of GNP is an equilibrium level?
6. Suppose that planned demand equals $400 billion and GNP equals $200 billion.
   a. Could the economy be in equilibrium? Why or why not?
   b. Explain carefully how the level of income would move to an equilibrium.
7. In your own words, explain why planned additions must equal planned withdrawals if planned demand = GNP.

25

The Multiplier

As you read and study this chapter, you will learn:

- why autonomous changes in planned demand usually trigger a multiplied expansion or contraction in GNP
- what determines the size of the multiplier involved
- when this multiplier process works, and when it does not
- why the multiplier is a key factor determining the cyclical stability of GNP

Suppose that an alien starship were to destroy Saturn and disperse its mass in outer space. The nicely balanced forces that hold the other planets in their orbits would be drastically altered, and here on earth we could expect quite a change in our weather. Or suppose that the Spellbinder cast his wicked magic on your favorite roller coaster, turning a swoop into a loop. Your anxiety about jumping the track might well turn out to be justified.

Newtonian mechanics could be used to calculate whether the roller coaster would leave the track, and whether the planets (minus Saturn) would crash together, fly apart, or settle into a new system of orbits. It could also calculate the paths they would follow. The first thing that would be needed for these calculations would be data on the masses and directional velocities of the planets and the Sun at the moment when Saturn was destroyed. The second would be the magnitude of the catastrophe—the dispersal of Saturn's mass. The third would be Newton's laws of

motion, which could be applied to convert the data into the predicted paths of the surviving planets.

Multiplier theory, which is the subject of the present chapter, is a lot like Newtonian theory as it might apply to the destruction of Saturn. It describes what happens to the circular flow of economic life when the behavior of one of the sectors changes spontaneously. In the language of multiplier theory, such a behavioral change is called an ***autonomous shift*** in demand. It is the equivalent of Saturn's sudden disappearance. Multiplier analysis also includes ***induced changes,*** the reactions that spread the effects of autonomous shifts and multiply them. These are like the rearrangements of the planets. If there were a drastic increase in military expenditures, for example, the full effects of the increase could be calculated by applying multiplier analysis. The raw material for such a calculation would consist of three parts. The first would describe the circular flow before the increase in military spending. The second would be the magnitude of the change. The third would comprise the laws of behavior of the sectors of the economy and a clear picture of the pattern of connections among them. Since you have already mastered most of this material in learning about the circular flow and its equilibrium, this chapter will not contain many new concepts. It will just combine familiar information in new patterns.

The multiplier was first incorporated systematically into macroeconomic theory in the 1930s, when Keynes made it the cornerstone of his analysis of the Great Depression. The concept has since been refined and quantified, so that it is now one of the most useful concepts economists have for studying short-run changes in income and output.

When an economist speaks of the "multiplier," he or she may be referring to a theorem, a process, or a number. The *multiplier theorem* states that in a market economy, any autonomous change in real planned demand leads to a cumulative reaction in the equilibrium level of production that is some multiple of the autonomous change that gets it started. The *multiplier process* is the working out of this cumulative response through a definite sequence of actions and reactions among the sectors of the circular flow. The total of the cumulative reaction in production relative to the autonomous change in demand is the *multiplier number.*

It goes without saying that the "multiple" is not one for one. If it were, the theorem wouldn't be very interesting. One of the things that makes the multiplier the centerpiece of macroeconomic theory is that it magnifies small autonomous changes into larger fluctuations in GNP. Economists draw two lessons from this. The first is that what looks like a general economic collapse may, in fact, have a specific and localized origin. Even though all sectors of the economy are in trouble, the source of their common problem may well lie in the particular difficulties of a single major industry. This sharpens the theory of economic fluctuations considerably, since it helps to pinpoint the sources of instability. The second lesson suggests a remedy for instability. If the government can limit autonomous fluctuations in demand, or offset them with opposing fluctuations elsewhere in the economy, then comparatively small government actions can prevent widespread instability in output, income, and employment. This particular lesson of multiplier theory accounts for its appeal to economists who think the government can and ought to take an active part in correcting what they see as faults of the market economy.

This chapter is organized into three sections. The first explains the multiplier theorem. The second describes the multiplier process and how it is linked to the size of the multiplier number. The final

section focuses on the *uses and limitations of the multiplier.* This section is a natural introduction to the theory of inflation and a necessary preface to studying the macroeconomic effects of the government budget. This study is one of the two main branches of the theory of ***stabilization policy,*** which consists of deliberate government efforts to control the business cycle.

The multiplier theorem

If you have mastered the equilibrium lessons of the preceding chapter, you have already grasped one of the central macroeconomic concepts. But it is just as important to understand how GNP moves from one equilibrium to another, and what factors determine the size of the movement. This is what multiplier analysis will teach you.

Clear thinking about how economic systems work requires clear concepts. ***In analyzing change, it is essential to distinguish between those movements that are imposed on the system from outside and those that result from the working out of the system's own internal laws. The first of these two kinds of change is called autonomous. The second kind is called induced.*** An autonomous change is a change in behavior, a change in the properties of the system. An induced change is the expression of established behavior, a working out of the properties of the system as it already exists. The distinction between these two kinds of change is particularly important for understanding the multiplier.

The multiplier in action

Rather than having you work your way through a lot of preliminaries, we will turn directly to an example of the multiplier in action. After you have seen it at work, you can study more carefully *why* it works.

Table 1 GNP and planned demand in a two-sector economy

GNP	C	I	Planned Demand	UI
600	600	40	640	-40
700	680	40	720	-20
800	760	40	800	0
900	840	40	880	20
1000	920	40	960	40
Equilibrium GNP—goes from 800 to 900			When planned investment rises from 40 to 60	
600	600	60	660	-60
700	680	60	740	-40
800	760	60	820	-20
900	840	60	900	0
1000	920	60	980	+20

Because of the dependence of consumption on GNY, equilibrium income and output must rise by 100 when planned investment rises by 20. The additional 80 is necessary to supply the induced increase in consumption.

The top half of Table 1 presents some figures that you encountered in Figure 4 in the last chapter. A level of planned investment equal to 40 goes with an equilibrium GNP of 800, since that is the only level of GNP at which production and planned demand are equal.

Now look at the bottom half of Table 1. Planned investment is 60 rather than 40. This change from 40 to 60 is an autonomous shift. It is not itself the result of a change in GNP. The consumption function that lies behind the figures at the bottom of the table is identical to the one that lies behind those at the top. You can prove this by noting that the level of consumption corresponding to each level of GNP is the same at the bottom as it is at the top. Thus, the only effect of the autonomous shift in planned investment is to raise total planned demand by 20 at each level of GNP.

What happens, therefore, to equilibrium GNP? Does it also rise by 20? If you look at the two halves of the table, you can see that in fact it rises by 100! An *autonomous shift* of 20 *induces an additional increase* of 80, leading to an overall change that is five times the size of the autonomous shift. This factor of 5 is the multiplier built into Table 1.

Although it may not be obvious, the multiplier also works in reverse. An autonomous decrease in demand is also amplified by induced changes. To see that this must be so, go back to Table 1, and think of the bottom half as "before" and the top half as "after." An autonomous drop in planned investment from 60 to 40 results in a multiplied drop of 100 in equilibrium GNP, since there is an induced drop of 80 in consumption.

Why does the multiplier work?

Magicians have only one principle: When you stuff the rabbit into your hat, make sure that everyone in the audience is looking somewhere else. If you look again at Table 1, you will probably see that a multiplier rabbit somehow got into an equilibrium hat. A change of 20 in planned demand produced a change of 100 in equilibrium GNP. Of course, you were expecting this, since it was printed right there on the advertisement at the beginning of the chapter. The only question is which distraction made it possible to put the rabbit into the hat without your seeing it.

In fact, it got there in the last chapter. The consumption function is responsible for the existence of the multiplier. When GNP rises to meet an autonomous demand increase, the higher production level generates more disposable income. This increase in household income leads to higher consumption demand. ***If GNP rises only by the amount of the autonomous demand increase, there will be unplanned disinvestment in inventories of consumer goods, since the higher GNP will induce an increase in consumption. The total increase in equilibrium GNP will have to be large enough to cover both the autonomous and the induced increases in planned demand.*** That is why the multiplier must be larger than 1. A one-for-one matching of additional GNP and additional autonomous demand won't produce equilibrium, since it won't cover the induced demand increase.

Of course, if there were no induced change in demand, the multiplier would be 1. Suppose that planned demand were entirely autonomous and completely independent of GNP. If planned demand were to increase autonomously, a matching increase in GNP would exactly cover the additional demand. With no induced increase in consumption, the total demand increase would equal the autonomous increase. No additional goods would be needed to supply induced consumption demand.

Many students find it helpful to study the multiplier with a diagram. The most important part of the diagram is the ***planned demand schedule,*** which shows what planned demand will be at various levels of GNP.

It is a lot like a consumption function, but it covers all planned demand, not just consumption. In the simple two-sector example of Table 1, however, planned demand has just two parts. First, *consumption* depends in a direct, uncomplicated way on the levels of GNP, GNY, and disposable income, all of which are equal because of the absence of taxation and business saving. This dependence, of course, is just the consumption function. Second, *planned investment* is a fixed amount, independent of the level of GNP. Since the government and foreign sectors are ignored until the next section, these two components make up total planned demand. It may be expressed as a fixed

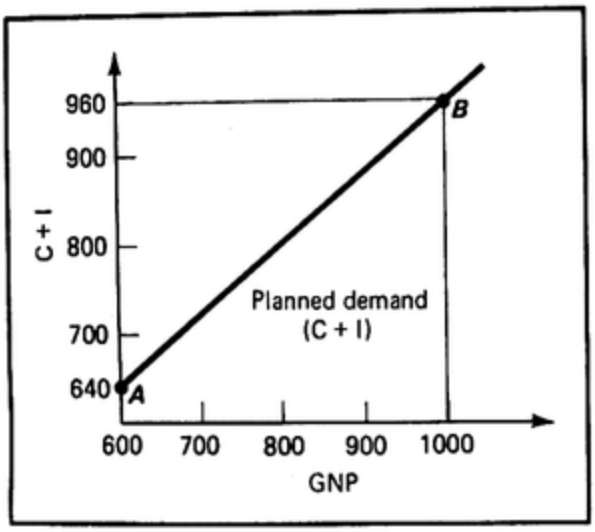


Figure 1 The planned demand schedule

This planned demand schedule shows what planned demand will be at every level of GNP. It corresponds to the top half of Table 1. For example, at a level of GNP equal to 600 (Point *A*), planned demand is 640. At a GNP of 1,000 (Point *B*), planned demand is 960.

amount (investment) plus a variable amount (consumption), dependent on GNP. If you look at Figure 1, you will find a planned demand schedule corresponding to the top half of Table 1. You can verify that the schedule in Figure 1 corresponds to the table.

This diagram makes it easy to envision what is meant by an autonomous shift in demand, and what is meant by an induced increase. Look at Figure 2, which illustrates both autonomous and induced changes. There are two planned demand schedules, *DD* and *D'D'*. Suppose an economy is initially at Point *A* on schedule *DD*. If planned demand increases autonomously, it will move to *B*, on *D'D'*. There is a shift in the demand schedule, from *DD* to *D'D'*, and a resulting change in demand, from *A* to *B*. But there is no change in GNP, which remains fixed at GNP_0. Now suppose GNP rises from GNP_0 to GNP_1, with no further autonomous increase in demand. This is represented by a *movement along* the new planned demand schedule, from *B* to *E*. The GNP goes up, and demand rises in response, with no change in anyone's behavior pattern. Autonomous shifts are *changes in behavior*, represented by shifts in the schedule. Induced changes are *consistent with established behavior* and are represented by movements along the schedules.

The planned demand schedule in Figure 3 illustrates the multiplier. There are three lines in the diagram. Two of them are planned demand schedules, *DD* and *D'D'*, corresponding, respectively, to the top and bottom halves of Table 1. The third line is the "line of equality." Since it has a slope of 45 degrees, and since the two axes have the same units, planned demand and GNP are equal at every point on this line. Obviously, every equilibrium must lie on the line of equality. Everywhere above this line, planned demand is bigger than GNP. Everywhere below, it is smaller. But since equilibrium requires that planned

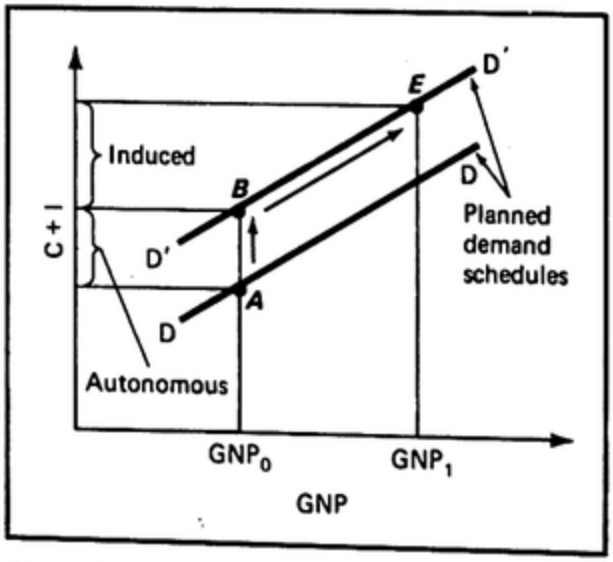


Figure 2 Autonomous and induced increases in planned demand

An autonomous shift in planned demand is a shift in the schedule, at a fixed level of GNP. The movement from *A* to *B* is autonomous. But an induced increase requires a change in GNP. The movement from *B* to *E* is induced. Notice that in going from *A* to *B*, demand changes as a result of a shift in the schedule, with *no change* in GNP. In going from *B* to *E*, it moves along a *given* schedule, in response to a *change* in GNP.

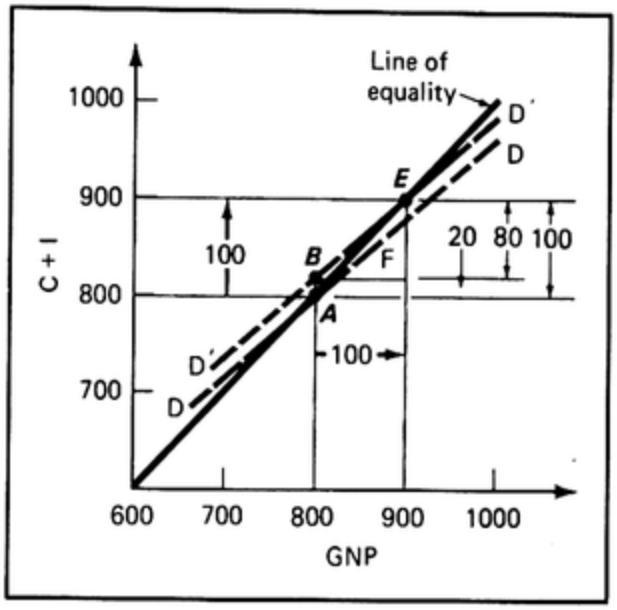


Figure 3 Equilibrium and the multiplier

This diagram corresponds to Table 1. The planned demand schedule *DD* matches up with that in the top half of the table; *D'D'* matches up with that in the bottom half. Each has a slope of 0.8, since the marginal propensity to consume (MPC) is 0.8 and investment is autonomous, independent of GNP. The difference between *DD* and *D'D'* is a constant vertical amount equal to 20, the difference between the two levels of planned investment (60 and 40). This is the distance from *A* to *B*. But the intersection of *DD* and the "line of equality" (at which planned demand equals production) is Point *A*, at a GNP level of 800. The intersection of *D'D'* and the line of equality is Point *E*, at a GNP level of 900. A difference of 20 in planned investment results in a multiplied difference of 100 in the equilibrium level of GNP. The reason for the multiplication is the induced increase in consumption, which rises by 80 for every 100 increase in GNP.

demand and GNP be equal, there cannot be an equilibrium anywhere but on the line of equality. Thus, a change in GNP without an autonomous change in planned demand cannot change the equilibrium. The equilibrium can only change when the schedule of demand intersects the line of equality at a different point.

Suppose that there is a shift in the planned demand schedule. Figure 3 shows what happens when it shifts upward by 20, from *DD* to *D'D'*: The equilibrium GNP changes by 100, from 800 to 900. The two equilibrium points, which are represented by *A* and *E*, are far apart relative to the magnitude of the autonomous change. The reason for this multiplication is the induced increase of 80 in consumption, which rises by 80 for every 100 increase in GNP. This means that an increase of 100 in GNP leaves exactly "enough room" for the autonomous increase of 20 in planned investment and the induced increase of 80 in consumption.

Recall from the last section that the multiplier works on both the downside and the upside. You can see this in Figure 3 by supposing that the planned demand schedule is initially *D'D'* and that it drops to *DD*. The autonomous fall of 20 in planned investment is represented by the movement from *E* to *F* at a fixed GNP of 900. But this induces a further decrease in planned demand, as falling GNP, GNY, and *YD* lead households to consume less. As a result, GNP drops from 900 to 800 overall, and the equilibrium is located by sliding down the new planned demand schedule, *DD*, to where it intersects the line of equality.

The multiplier process

The example and the multiplier number worked out in Table 1 show *why* the multiplier must exist. An autonomous demand increase causes a production increase, which then results in a further demand increase. The total production increase must be big enough to supply both the autonomous and induced increases in demand. This also works in reverse. If demand drops autonomously, the resulting drop in production will cause a further drop in demand. The overall production cut must match both the autonomous and the induced cutbacks in demand.

Even though the *logic* of the multiplier theorem is overwhelming, it doesn't say enough about *how* these changes occur. A

fully convincing account of the multiplier must also outline the process of change. Why does it work? What do people do collectively that makes it work?

Actions and reactions

The multiplier process is rooted in the circular flow and grows out of the interdependence among the sectors. Suppose, for example, that American automobile manufacturers develop better cars than their Japanese competitors. Consumer demand for cars may not change, but sales of imports will drop and domestic dealers will enjoy a corresponding increase in their sales.

The first result will be a change in dealer inventories. The diversion of consumer demand from imports to domestic cars will lower the inventories of GM dealers and lead to an unintended buildup of unsold Toyotas. So far, nothing in the GNP accounts has changed, but one part of the retail automobile sector has suffered an unintended inventory loss and another an unintended gain.

Of course, those dealers that lose inventories are the winners. Since their sales are larger, they will order more cars from their domestic suppliers. The Toyota dealers, with their larger stocks of unsold cars, will reduce their orders. More GM cars and fewer Toyotas will be seen on the trailers of automobile transporters.

It is at this stage that the shift in demand starts to affect GNP. Output and employment rise in Pontiac, Michigan, and fall in Toyota City, Japan. This shows up in the U.S. national accounts as a drop in imports, with no change in the other components of final demand. It also means a higher level of U.S. output, since the (unchanged) total sales of automobiles are now supplied by a higher flow of U.S.-produced cars. The reactions don't stop with the automobile industry. They are spread throughout the economy by several channels. The most direct is the input-output network. The automobile companies buy steel, glass, tires, wire, paint, plastics, carpet, mufflers, radios, nuts and bolts, and an incredible variety of things from other companies. Since many of these other companies specialize in supplying the auto industry, their fortunes rise and fall with those of the auto companies. If GM produces more cars, these firms will produce more steel, glass, tires, and so on, and employ more workers. Thus, production will go up throughout all the far-flung industries that directly and indirectly add their value to the value of finished automobiles, and GNP will rise.

Another channel, almost as direct, is how households react to higher employment, wages, and profit income. Both in the auto industry and in its various suppliers, the increased sales of domestic cars will mean increased domestic income for workers and owners of capital. Part of this additional income will be spent on consumer goods, spreading the impact of prosperity to sectors of the economy far removed from the production of automobiles. The input-output relations of these sectors will spread the reaction still further.

A third channel, less direct, is that the various sectors of the economy that share in the increased prosperity may decide to increase their productive capacity. They will invest in plant and equipment that they otherwise would not have needed. This will raise production in the investment goods industry, and indirectly in all those industries that supply it with intermediate goods. It will also raise the incomes—and consumption—of those households that supply factor services to the investment goods and related industries.

There are three offsets to this spread of domestic prosperity. First, some part of the rising income will go to pay taxes, and another part will be saved. Both will

limit the expansion of demand and production. Second, since many of the inputs to the increased production will come from abroad, imports will rise, diverting part of the increased demand to other countries. Some of the higher U.S. income will even be spent on Toyotas, so that the total loss in Toyota sales will be smaller than the autonomous drop. Third, since Japan will earn fewer dollars by selling to the United States, either Japan or one of its other trading partners will buy less from us.

The MDP

The discussion of the previous section intentionally omitted any specific numbers. It sought to make you concentrate on the economic processes and motives that are responsible for multiplier expansion. But the multiplier can also be given precise quantitative expression in terms of the behavior of the sectors that make up the circular flow.

Although it would be possible to work through the multiplier process in terms of the simple two-sector model, there is no point in starting at this level. You are now familiar enough with the four-sector economy to follow it through. It is not really more complicated, just a bit more cluttered.

The clutter can be reduced by introducing yet another concept, the ***marginal demand propensity.*** It is new to you only in name. The planned demand function has a slope, which shows how much planned demand increases per unit increase in GNP. This slope is the marginal demand propensity, or *MDP*.

Like the marginal propensity to consume, the marginal demand propensity shows how one thing changes in *response* to a change in something else. But the MDP is broader in its coverage than the MPC. It measures how *total planned demand* responds to a change in GNP. This means that it summarizes the behavior of all the sectors. This behavior is entirely *induced.* Autonomous changes are represented by shifts in the planned demand schedule, not by its slope.

Because it is the slope of the planned demand function, the MDP can be written as:

$$MDP = \frac{\text{Change in } (C + I + G + X - M)}{\text{Change in } GNP}$$

$$= \frac{\Delta C + \Delta I + \Delta G + \Delta X - \Delta M}{\Delta GNP}.$$

In the two-sector example of Table 1 and Figure 3, the slope of the planned demand schedule was 0.8. This was, of course, the MPC because I was treated as entirely autonomous, G, X, and M were ignored, and no taxes or business saving intervened between GNY and disposable income.

In fact, things are more complicated than that. At the margin, business saving absorbs about 12 percent of GNY. Net taxes absorb about 25 percent. Since these two leakages collectively absorb 37 percent, the change in disposable income is only about 63 cents on every dollar change in GNP and GNY. If the MPC out of disposable income is 0.8, then $\Delta C/\Delta GNP$ is about 0.5 ($0.63 \times 0.8 = 0.504$).

Government purchases *(G)* and exports *(X)* are largely autonomous, but a change in GNP induces changes in investment *(I)* and imports *(M)*. You may recall from the last chapter that the size of $\Delta I/\Delta GNP$ is a matter of controversy. The examples in that chapter used the figure 0.15, and this chapter will do the same. As for imports, the value of $\Delta M/\Delta GNP$ is also about 0.15, but imports enter into final demand with a negative sign. So:

$$MDP = \frac{\Delta C}{\Delta GNP} + \frac{\Delta I}{\Delta GNP} + \frac{\Delta G}{\Delta GNP} + \frac{\Delta X}{\Delta GNP} - \frac{\Delta M}{\Delta GNP}$$

$$= 0.5 + 0.15 + 0 + 0 - 0.15$$

$$= 0.5.$$

The figures that make up this MDP are about the right size for the U.S. economy, although there is a substantial margin of error in measurements of this kind. Whatever the margin of error, the MDP is close to 0.5, not to zero or to 1.

A more realistic example

To see how the MDP enters the multiplier process, imagine an autonomous increase in planned demand. Suppose that it is a rise in government purchases of $1 billion. If the economic system gears up to produce an additional $1 billion of GNP per year, it will supply the additional autonomous demand. But the structure of the circular flow will not let the process stop there. The $1 billion additional GNP will generate additional income and additional demand. If the MDP is 0.5, the $1 billion rise in GNP will induce an additional $0.5 billion in demand. If goods are produced to supply this demand, this will generate still more income, and still more demand.

Where will this process end? If you have ever worked your way through geometric series, you can probably figure out the answer. Look at the following sequence of numbers:

1. First-round *autonomous* demand increase = $1 billion.
2. Second-round *induced* demand increase = $0.5 billion.
3. Third-round *induced* demand increase = $0.25 billion.
4. Fourth-round *induced* demand increase = $0.125 billion, etc.

You can, if you like, carry this table on forever. Each term is half as big as the preceding one, because the MDP is 0.5. Eventually they get very small. But what happens to GNP depends on the *sum* of the successive reactions. The arithmetic of geometric series tells you that even though the series has an infinity of terms, it has a finite sum, equal to 2. It is equal to $\frac{1}{1 - MDP}$. (There is no particular mystery to this formula. It comes from the algebra of geometric series.)

You can also work through the same arithmetic for an autonomous *drop* of $1 billion in government purchases. As a result of such a drop, GNP will be cut back by $1 billion, consumption will fall by $0.5 billion, planned investment and imports will each drop by $0.15 billion. Thus, the second-round induced drop in demand will be $0.5 billion. But the process will not stop with the second round. There will be a succession of induced rounds of GNP contraction, each half as big as the one that preceded it. Altogether, the induced changes will add up to a negative $1 billion, for an overall drop of autonomous plus induced demand equal to $2 billion. The autonomous negative $1 billion change in planned demand will be doubled by the induced effects, just as the autonomous positive change was doubled in the previous example.

The MDP and the size of the multiplier

The size of the MDP determines the size of the multiplier. The precise relationship is easiest to see by looking at some fairly simple algebra. Suppose there is an autonomous demand increase equal to ΔA. The overall demand increase will equal the autonomous increase plus the induced increase. The induced increase will equal the product of the MDP and the overall increase in GNP. If this overall increase is written as ΔGNP, the relationship between ΔGNP and ΔA can be found from

$$\underset{\text{overall increase}}{\Delta GNP} = \underset{\text{autonomous increase}}{\Delta A} + \underset{\text{induced increase}}{MDP \times \Delta GNP}$$

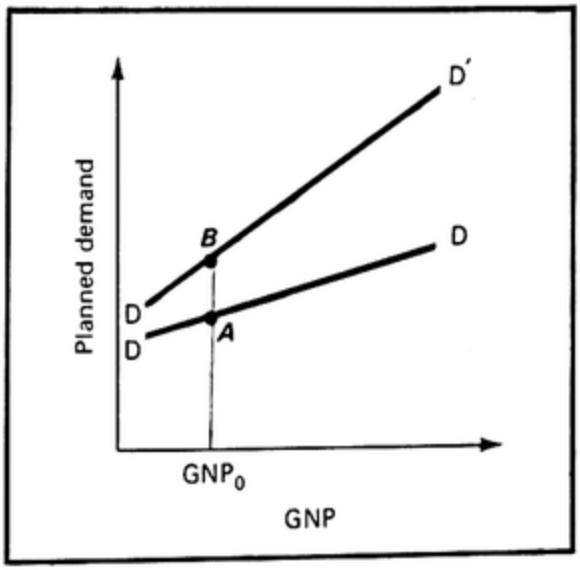
(equals, plus)

If the second term on the right is moved to the left and GNP is factored out separately, the equality reads:

On Shifts and Slopes: Taxes, Transfers, and Planned Demand

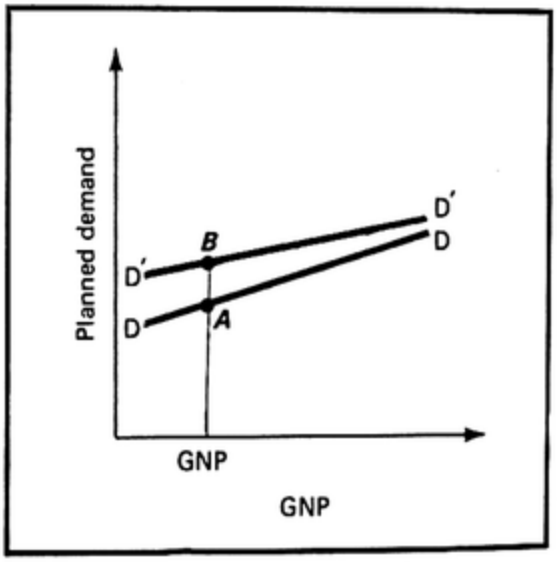
Every form of spending behavior and every aspect of the distribution of income influences the planned demand schedule. This is true both for the government and for the private economy. When you study fiscal policy in later chapters, you will learn the major ways in which governments, especially the federal government, influence the level of GNP. For now, it will enrich your understanding of the multiplier to see where the government enters into the picture.

Take taxes: High tax rates absorb large amounts of income and reduce planned demand. Low tax rates do the opposite. If the government reduces its tax rates, it raises planned demand, since it leaves more GNY in the hands of consumers and raises *C* relative to GNY. Reducing tax rates also raises the slope of the planned demand schedule. If tax rates are lower, the government takes a smaller bite of GNY at the margin. This leaves more of an increase in income in the hands of consumers and lowers the government's share of a decreased income. The result is to raise the change in disposable income relative to a change in GNY. In effect, the marginal propensity to consume is raised, not because people spend more of a change in their disposable income, but because *YD* changes more for every



Effects of a tax rate reduction

A cut in income tax rates raises consumption and therefore planned demand at every level of GNP. For example, it might rise from *A* to *B* at a level of GNP equal to GNP_0. But the slope also goes up. Since taxes take a smaller bite out of increases in GNY, a bigger share is left over, and both the effective MPC and the slope of the planned demand schedule are bigger.



Effects of an increase in unemployment compensation

An increase in the amounts paid to the unemployed also raises the planned demand schedule at every level of GNP. But this change lowers the slope of the planned demand schedule. As the unemployed go back to work when GNP rises, the amount that their income rises is smaller if the unemployment compensation they lose is larger. With a smaller rise in disposable income as GNY increases, both the effective MPC and the slope of the planned demand schedule are smaller.

dollar change in GNY. This means that a reduction in tax rates not only shifts the planned demand schedule upward, but also increases its slope!

The effect of an increase in unemployment compensation is quite different. This, too, shifts the planned demand schedule upward. The higher unemployment compensation is, the higher *YD* is relative to GNY. Less is drained off in taxes minus transfers, which include unemployment compensation. But notice that if you get a large unemployment check when you are out of work, you lose a lot when you go back to work. This means that the rise in disposable income per dollar rise in GNP is not as high as it would be if unemployment compensation were smaller. The increase in wages that goes with greater output and employment is offset by a big drop in unemployment compensation. Thus, consumption also changes less drastically when GNP and GNY change. Since the responsiveness of consumption is one of the main determinants of the slope of the planned demand schedule, its lower responsiveness means a less steep planned demand schedule.

$$\Delta GNP\ (1 - MDP) = \Delta A.$$

Solving this for $\Delta GNP/\Delta A$ gives:

$$\frac{\Delta GNP}{\Delta A} = \frac{1}{1 - MDP}.$$

Since this is the total increase in GNP divided by the autonomous increase that initiates it, it is the multiplier. Provided that the MDP is between zero and 1, this multiplier is greater than 1 but finite.

The multiplier and stability

To put the size of the MDP in perspective, think again of the multiplier process, of the successive rounds of demand expansion. For every dollar of autonomous increase in planned demand, there is a series of responses that add up to the corresponding increase in GNP, so

$$\frac{\Delta GNP}{\Delta A} = 1 + MDP + MDP^2 + MDP^3 + \ldots.$$

Notice that the bigger the MDP is, the bigger every term in this series past the first also is. In the language of the multiplier process, the bigger the MDP is, the bigger the successive rounds of response to an autonomous demand increase are, and the bigger the multiplier, their sum, is. If the MDP and therefore the multiplier are small, the response to autonomous demand fluctuations will remain within narrow bounds. But the bigger the successive rounds are, the more *unstable* the economy will be. Any economy that "takes off" in response to small autonomous fluctuations in demand will have wide swings in its business cycle.

In fact, if the MDP were greater than 1, the multiplier would not even be finite. You can see that if this were true, the successive rounds of expansion would not die out but "explode." The multiplier process would be totally unstable. Fortunately, this is not the case for the American economy. Careful numerical estimates of the MDP place it in the general neighborhood of 0.5, and the multiplier in the neighborhood of 2. Withdrawals into business and personal saving, net taxes, and imports greatly outweigh induced increases in investment, so that the successive rounds of the multiplier process die out rather rapidly.

Factors affecting the size of the multiplier

If the multiplier must be $1/(1 - MDP)$, a rise in the MDP must also raise the multiplier, since it will lower the denominator. But while true, all of this is mechanical. To be really comfortable with the multiplier, you need to see how its size is affected by changes in the behavior of the sectors of the economy.

The three demand components that are mainly responsible for induced demand changes are C, I, and M (which enters final demand with a negative sign). The values of G and X are largely autonomous in the short run.

Consumption changes reflect the habits and calculations of the household sector as it responds to changes in disposable income. Recall, however, that the MDP depends on how C responds to GNP, not just to YD. Both the business and government sectors intervene in this response, since business saving and net taxes are drained off from GNY before the remainder becomes YD. So the consumption response to GNP depends on how all three domestic sectors behave. The precise relationship is:

$$\frac{\Delta C}{\Delta GNP} = MPC \cdot \frac{\Delta YD}{\Delta GNP}$$

or

$$\frac{\Delta C}{\Delta GNP} = MPC \left(1 - \frac{\Delta SB}{\Delta GNP} - \frac{\Delta TN}{\Delta GNP}\right).$$

The induced change in consumption per dollar change in GNP equals the marginal propensity to consume (MPC) times the change in disposable income per dollar change in GNP. The latter equals 1 less the marginal share of GNY that leaks away into business saving and net taxes.

In this chapter, the figure that is used for $\Delta C/\Delta GNP$ is 0.5. Remember how this is arrived at: The MPC is about 0.8 in the short run. Business saving absorbs about 12 percent of GNY at the margin, and net taxes about 25 percent. Thus:

$$\frac{\Delta C}{\Delta GNP} = 0.8\ (1 - 0.12 - 0.25)$$
$$\frac{\Delta C}{\Delta GNP} = 0.8\ (0.63) = 0.504$$

or about 0.5.

Notice that any of the following things would *lower* $\Delta C/\Delta GNP$, the MDP, and the multiplier:

1. A drop in the MPC.
2. A rise in $\Delta SB/\Delta GNP$.
3. A rise in $\Delta TN/\Delta GNP$.

Obviously, a lower MPC would imply that consumption was less responsive to disposable income, and therefore to GNP. A rise in the share of GNY absorbed at the margin by either business saving or net taxes would insulate disposable income, and therefore consumption, from the effects of changes in GNP. The American economy has a small multiplier and a relatively high degree of stability because net taxes do so much of this insulation. You will learn more about this in a later chapter.

The figure of 0.15 for $\Delta I/\Delta GNP$ that is used in this chapter is largely arbitrary and illustrative. If a bigger figure had been used, the MDP and multiplier would have been bigger. The size of the investment response to changes in GNP is one of the major weaknesses in our knowledge of how the American economy works. You learned some of the reasons for this in the last chapter. Investment forecasts are uniformly rotten. Forecasters have to put up with a lot of uncertainty in this area, which leads to a corresponding uncertainty about both the multiplier and the likely autonomous changes in planned investment. If the actual response of I to changes in GNP is greater than 0.15, the multiplier is bigger than 0.5. If it is smaller, the multiplier is smaller too.

The figure of 0.15 that is used for $\Delta M/\Delta GNP$ is not arbitrary, but roughly consistent with experience. If this figure were bigger, the multiplier would be smaller.

Imports are a leakage from the circular flow. The more they respond to changes in production, the smaller the impact of GNP changes is on *domestic* production and income. This means that very "open" economies, with a great dependence on imports, have smaller multipliers than "closed" economies, with limited foreign trade. The American economy is much more closed than most of the other major industrial countries, partly because it is also a major agricultural producer.

The uses, misuses, and limits of the multiplier

When the multiplier first entered the economic literature in the work of Keynes during the 1930s, it was something of a puzzle. Economists weren't nearly so good at mathematics then as they are now, and even those who were (like Keynes) didn't understand the process behind the number very well. Once the multiplier process was understood, it moved in and took over much of the macroeconomic household. The mathematical economists in particular were intrigued with it, and many older economists without much mathematical training were pushed into the background. Accumulated wisdom about cumulative economic processes and their relationship to the stock of money got crowded out of the professional journals. The older generation lost its audience, the younger generation got the attention, and for two or three decades Keynesianism was the rigid orthodoxy in the profession.

Karl Marx once wrote that he was glad that he, at least, was not a Marxist. Keynes never said as much about himself in relation to his followers, but he probably did not fully sense the direction they were taking. In many ways, Keynes was a very traditional economist, well aware that capacity to produce always places an outside limit on GNP, and that the economics of output, employment, and the circular flow does not begin and end with the multiplier.

But the multiplier is not just a gimmick either. It is a very useful tool. First, it helps us understand why so many sectors of the economy go up and down together. Second, it can be used to make numerical estimates of the effects of changes in autonomous spending. Third, it can be used to estimate the quantitative government policy changes needed to smooth out some of the business cycle.

The multiplier and the whole approach to economics that it represents has recently been attacked by some very astute people. Much of the controversy could be resolved if people would always be careful to apply the multiplier theorem only where it works. For it has definite limitations.

To see why the multiplier theorem has to be used with care, contrast two situations. First suppose that excess capacity and unemployment are widespread among experienced workers in the U.S. economy. If the government, for example, increases its purchases of goods and service, this will put people to work and use more productive capacity. Through the channels of the input-output structure, the expansion will spread. Incomes, consumption, and investment will rise throughout the economy.

But suppose there is high capacity use and low unemployment. If the government raises its demand for goods and services, this cannot call forth increased production without creating bottlenecks and higher costs. Either government demand will be frustrated, or the resources necessary to produce government goods will have to be diverted from somewhere else, *crowding out the production* necessary to satisfy other demands. Until something happens to *crowd out the demands themselves*, de-

mand as a whole will be greater than the economy's immediate capacity to produce. What will happen? There will be shortages and inflationary pressure. Prices will rise. GNP in current dollars will go up, but real or constant-dollar GNP will not. To understand this, you must first see that when we are close to full employment, we are in *the territory of the inflationary process, not the territory of the multiplier process.* The boundary between these two territories is wide, blurry, and, for a great variety of reasons, shifting. And since the laws on one side of the boundary are quite different from the laws on the other, it pays to know where you are.

This crowding out may occur even when there *appear to be* sufficient unemployed workers and unused capacity to allow increases in output to take place. Think of the economy as having both a mainstream and a backwater. The mainstream work force is experienced and mature, predominantly white and largely male, although it includes some nonwhites and increasingly large numbers of women. These mainstream workers have in common a record of experience and regular attachment to the labor force. In contrast, the backwater labor force is made up of young people of all races and many nonwhites of all ages. These workers have little job training, limited experience, and very high unemployment rates. If they had more experience, they would not be a "backwater," and their ability to step into jobs and produce additional output would be real rather than apparent.

Suppose that the economy is moderately prosperous, but the unemployment and capacity utilization data tell us that there is plenty of room for expansion. Then an autonomous increase in demand occurs, and the multiplier process amplifies it and carries its effects into all corners of the economy. Unemployment in the mainstream labor force quickly reaches a minimum—and overtime a maximum—at what is usually called the *frictional level.* Although there is still unemployed labor, further expansion of output must draw heavily on teenagers, blacks, Hispanics, and women who are not regularly part of the labor force. Their inexperience would make it difficult to integrate them quickly into work processes, even if prejudice and discrimination did not add to the problem.

Something similar seems to happen with productive facilities. Once the modern capacity of established firms is in full operation, the only slack to be found is in new firms and in the outmoded "standby" capacity of older firms. The older facilities are inefficient, and newer suppliers may not easily be located by the market. As a result, capacity bottlenecks develop throughout the input-output structure. Thus, both the labor force and the capital stock may limit expansion, even when there are still unemployed workers and unused capacity.

When you think about unemployed resources and the multiplier, you should also think about mainstream resources. The extent and makeup of the mainstream can change, but only gradually. Both the development of an experienced work force and the expansion of plant capacity take time. They can and often do keep pace with moderate growth in demand. But because the time lag limits the speed at which supply can grow, it can also limit the applicability of the multiplier, even when unemployed resources exist.

Public confusion about the multiplier effects of the federal budget reached a peak during the "stagflation" of the late 1970s. You may remember the kind of statements that appeared in the press around that time. Perennial critics of government spending argued that events were proving exactly what they had always believed:

"Government spending just causes inflation, it can't create jobs." Liberals pointed to the high and growing unemployment in the backwaters of the labor force and lobbied for more "fiscal stimulus" to expand demand and create jobs.

Both groups were confused; both were partly right but partly wrong. The conservatives were probably right in thinking that at the time, a reduction in government spending would have reduced inflation without adding much to unemployment. But they were wrong in stating categorically that increased government spending is always inflationary, never expansionary. The liberals were probably wrong in expecting much employment effect from higher government expenditure, and they probably underestimated the likely inflationary consequences. They were right, however, in recognizing that the budget can directly influence real income and employment. Their mistake was to try to apply multiplier thinking to the inflationary process.

These issues will be more thoroughly examined in the next chapter, and the 1970s in particular in a later chapter, which deals with government attempts to stabilize the economy. But before you can study this subject from an informed, analytical perspective, you will have to master some more analytical tools.

Summary

This chapter is short but difficult. It has been exclusively concerned with a key analytical concept known as the multiplier. Economists use this term to refer to a theorem, a process, and a number. The multiplier theorem states that in market economies, any autonomous change in real demand leads to a cumulative reaction in production, which is some multiple of the autonomous change. The multiplier process is the working out of this cumulative response through a definite sequence of actions and reactions among the sectors of the circular flow. The multiplier number, which is larger than 1, is the total of the cumulative reaction in production relative to the autonomous change in demand.

The multiplier works because when increased demand is met by increased production, the higher income that is generated induces a further demand increase. The size of this induced change in demand depends on the marginal responses of consumption, investment, and imports to changes in GNP and GNY. The sum of these responses relative to the change in GNP is called the marginal demand propensity, or MDP. For year-to-year changes, the MDP is about 0.5 in the U.S. economy. The multiplier number is equal to $1/(1 - MDP)$, so that it is about equal to 2.

Although the multiplier is one of the most useful tools that economists have for analyzing short-run changes in GNP, it has definite limitations. In particular, it doesn't work when there are shortages of experienced labor and productive capacity. When production is limited by such shortages, we are in the territory of the inflationary process, not the territory of the multiplier process. These two territories are ruled by quite different laws. The next chapter analyzes the inflationary process.

Key concepts

- Multiplier theorem, process, and number
- Planned demand function
- Autonomous and induced changes in demand
- Marginal demand propensity (MDP)

Questions for review

1. Indicate whether the following changes in planned demand are *autonomous* (a shift in the planned demand schedule) or *induced* (movement along the planned demand schedule). If there is an autonomous change, indicate whether there will be an upward or downward shift. Indicate those cases in which you feel there is insufficient information to make a judgment about the autonomous or induced nature of the change.
   a. Interest rates decrease and investment rises.
   b. Tax rates are lowered.
   c. The price of U.S. goods rises faster than the price of goods manufactured abroad, so that exports fall and imports rise.
   d. Tax revenues increase.
   e. Consumers increase saving.
   f. The government decreases its purchases of goods and services.
   g. Consumers expect a severe recession and reduce consumption.
   h. The level of exports decreases.
   i. Consumption rises because disposable income increases.
2. a. What kinds of changes will start the multiplier process?
   b. Suppose that government purchases are reduced by $800 million. Carefully and clearly explain the multiplier process, by which this $800 million cut is multiplied into an even larger decrease in equilibrium income. (Make sure that your answer includes a clear statement of why the multiplier process will end.)
3. Explain whether the following changes will increase or decrease the size of the multiplier, or leave it unaffected.
   a. A change in exchange rates makes foreign goods more expensive for U.S. citizens.
   b. Inheritance taxes are increased.
   c. Consumers fear that a recession will occur.
   d. Investment increases.
   e. Unemployment compensation payments per recipient increase.
   f. An increase in GNP leads to an increase in net taxes.
   g. Government reduces spending by $2 billion.
   h. There is a legislated increase in tax rates.
4. Explain the general relationship between "built-in stabilizers" (such as tax rates and transfer payments that vary with income) and the size of the multiplier.
5. a. When is the economy in multiplier territory?
   b. Suppose that data show that there are unemployed workers and unused capacity in the economy. Does this mean that the economy is necessarily in multiplier territory? Explain.
6. Two of your friends are having a heated argument. One claims that government spending must always be inflationary. The other claims that government increases output and not prices, resulting in significant decreases in unemployment. Use all of your economic knowledge to mediate this argument.

26

Unemployment and Inflation

As you read and study this chapter, you will learn:

- what is meant by inflation
- how changing costs affect prices throughout the economy
- why wages rise when unemployment is low
- why productivity gains are important for the stability of prices
- how demand changes affect the rate of inflation
- why persistent inflation is qualitatively different from temporary inflation

Most of us who have lived through the past decade have some very definite views about inflation. The inflation that began in the late 1960s has been the topic of the day in the news media, in the corridors of power, and at cocktail parties. Unlike the spectacular "hyperinflations" of Europe in the 1920s, when annual inflation rates were measured in billions, the United States has never produced rates of price increase that could be called socially disastrous. Yet, the increases of 5 to 12 percent that have followed one another year after year were unheard of in living memory. Rates that are historically high for a particular country generate uncertainty, instability, and anger. And you know from your own or your family's experience that inflation can be painful.

Dictionaries define inflation as a persistent rise in prices caused by excessive expansion in paper money and credit. So, if you think that inflation is a monetary problem, you are in good company. There are two objections to the dictionary definition,

however. First, it mixes a statement of what inflation is with a statement about why inflation happens. This is like defining death as an end of all normal metabolism caused by heart failure—what should we call the result of decapitation? Second, the indicated cause is very one-sided and incomplete. Excessive monetary growth may be a *precondition* for continuing inflation, but the *immediate causes* are far more numerous than the dictionaries lead you to believe.

A careful study of inflation must begin with a definition that does not assume a simple cause. This rules out the dictionary definition: It is best to use the term *inflation* to apply to *any rise* in the general level of the prices. Inflation may be said to be mild or rampant, brief or extended. There can be inflation in retail prices at the same time as stability or even decline (deflation) in producer prices. It is, however, useful to have a specific name for an inflation that is widespread and long-lasting enough to become the normal state of affairs. A good name for this is a ***persistent inflation.***

As soon as you think of prices, you probably think of supply and demand. The analysis of inflation, you will be glad to hear, is organized around supply and demand. But in studying the *general* price level with the tools of supply and demand, it would be foolish to forget the lessons of the last three chapters. Because of the structure of the circular flow and the input-output network, supply and demand changes are interdependent for the economy as a whole, in a way they are not for a single market.

The chapter is organized into four main parts. The first part begins with supply side and focuses on the cost of inputs. Inflation that results from rising input costs is usually called ***cost-push inflation.*** Labor, the major input to the production process as a whole, receives special attention. The second part, still focusing on supply, shows how productivity gains may offset the effects of input cost-push. The third section turns to the demand side. Inflation that originates in demand expansion is usually called ***demand-pull inflation.*** Of course, demand pull cannot be understood without understanding the multiplier process, which we discussed in the last chapter. Finally, the fourth section is devoted to the distinctive features of persistent inflation.

The cost of inputs

The prices of most products are fairly closely related to their costs of production. An increase in cost raises the prices that competitive firms must charge if they are to stay in business. In less competitive industries, a conspicuous rise in costs often stimulates producers to raise their prices jointly and to protect their profit margins. Almost the only products whose prices are not cost related are those in fixed supply, such as works of art, and those whose supplies are limited by powerful monopolies, such as crude oil.

Production costs vary directly with the prices of inputs. Although most producers can use some substitutes for inputs whose prices are rising, this substitution can be very limited in the short run. If you raise orchids in an oil-heated hothouse, you are just stuck when the price of oil goes up.

The cost of intermediate goods

To get some feel for the structure of costs in a representative manufacturing industry, look at Table 1. It shows the breakdown of the value of a typical dollar's worth of household appliances (TVs, stoves, washers, etc.) in 1972, as recorded by the 1972 input-output study. As you can see, each dollar's worth of output con-

Table 1 ***Allocation of value of household appliances among costs and profit, 1972***

Input	Cents of Cost or Profit per Dollar of Output
Paperboard containers and boxes	2
Rubber and miscellaneous plastics	4
Primary iron and steel	8
Primary nonferrous metals	6
Screw machine products and stampings	3
Other fabricated metal products	3
Electrical industrial equipment	5
Scientific and controlling instruments	4
Wholesale and retail trade	4
Business services	4
Other intermediate goods and services	15
Total intermediate goods and services	58
Employee compensation	27
Property-type income (depreciation, rent, interest, profits)	15
	100

In 1972, the cost of intermediate inputs was about 58 percent of the value of household appliances, labor cost was 27 percent, and depreciation, rent, interest, and profit costs were 15 percent.

Source: U.S. Department of Commerce, *Survey of Current Business.*

tained about 58 cents worth of intermediate goods and services—mostly materials, but also many miscellaneous services like lawyers' fees, telephones, insurance, and wholesale margins on and transportation of materials.

Besides the cost of the intermediate goods, a dollar's worth of appliances contained about 27 cents of employee compensation—wages, salaries, fringe benefits, and social insurance taxes paid by employers. The remaining 15 cents, which the Commerce Department labels "property-type income," was made up of depreciation, interest, rent, and profit. Indirect taxes were negligible.

Together, the employee compensation and property income added up to the value of production taking place in the industry, the *value added.* For the appliance industry, value added was 42 percent of the value of sales. The other 58 percent was the production of other industries, incorporated as inputs in the manufacture of appliances. In general, these figures are representative of manufacturing, utilities, and transportation. However, in agriculture, mining, and service industries, value added is a much bigger share of the total. In wholesale and retail trade, value added is a relatively small fraction of price.

In thinking about the relationship between prices and costs, it also helps to think about what happened to the appliance industry's output. Of the industry's roughly $7 billion output in 1972, about $1 billion went to other industries as intermediate inputs, mostly to the transportation equipment, repair, and construction industries. Another $1.4 billion went into capital expenditures of the hotel, motel, and rental housing industries. Both these amounts, marked up slightly for transportation, fed back into the economy's cost structure. The rest of the industry's output went to consumers. But the $4.6 billion of

sales by appliance producers was marked up to $8 billion before it reached consumers. Most of the markup was the retail trade margin, which covered the costs—wages, rent, utilities, interest—and profit of the retailers.

Given this picture of the flow of goods through the appliance industry, two facts stand out immediately:

1. The costs of any one industry are directly affected by what happens in industries *upstream* in the input-output structure—in its *supplier* industries. For the appliance industry, iron and steel, nonferrous metals, and electrical equipment stand out, but 7 other industries appear explicitly in Table 1, and the "other intermediate" grouping is an aggregate of 42 industries, each contributing less than 2 cents per dollar. Anything that affects the price of upstream products (including prices of goods still further upstream) will change the costs of appliance production.

2. Even a product category like appliances, seemingly made up of consumer goods, affects costs *downstream* in the production process—in its *user* industries. About a third of the industry's output became either an intermediate good, directly entering into the cost of recreational vehicles, aircraft, and so on, or a capital good, indirectly entering the depreciation expenses of hotel, motel, and rental housing owners. Thus, anything that affects appliance prices raises costs in these other industries, as well as the cost to retailers of the consumer appliances they sell.

When you studied the input-output accounts in an earlier chapter, you learned that the value of total output could be traced back from final demand to the production of the many industries that contribute to its creation and that production equals the sum of incomes. Given these two facts, the cost of any product, if traced far enough upstream, can be entirely assigned to employee compensation, various indirect taxes, rent, interest, depreciation, and profit. ***The intermediate goods are not an independent cost category, since their value can be traced entirely to wages, taxes, and property incomes along the way.*** This is one big truth about the structure of the economy. However, it hides another big truth: ***Intermediate goods are a major transmission route for spreading price changes throughout the economy.*** It is a two-way route. For example, if the demand for steel goes up and the steel industry has excess capacity, its output will rise. This will raise the steel industry's demand for inputs that originate upstream in the input-output flow. Even if the increase in the demand for steel doesn't directly raise the price of steel, its transmission to upstream suppliers may raise prices in those industries. These price increases in supplying industries will be passed back downstream to steel itself, in the form of increased input costs. If they are passed on by the steel industry, they will raise costs in every industry that directly or indirectly uses steel as an input, which is to say almost every industry you can think of.

Imports

Imports are netted out in the calculation of GNP, since their value is not part of domestic product. But their prices do affect domestic costs. If a firm's raw material prices have just gone up, does it matter whether those materials come from domestic suppliers or from abroad? Of course not. No matter where the inputs come from, if they cost more, they will push up the prices of the firm's output.

However, from the point of view of the society's real income, it does matter whether the materials are domestic or im-

ported. Think what happens if the price of a single domestic product goes up and all other prices stay the same. This reduces the real income of all users of the product and raises the real income of at least some of its producers. There is no direct effect on the GNP; it is merely redistributed. But suppose that the product whose price has risen is an import. The producers live abroad, but the users live here. Hence real income is redistributed to foreigners. Real GNP doesn't fall. But the real exports that must be sold to pay for the relatively higher priced imports go up, and domestic consumption must fall. This is a real cost.

The relationship among import prices, the domestic price level, and real income was brought home to almost every American by the great oil price rise of the 1970s. In 1973, the Organization of Petroleum Exporting Countries ***(OPEC)*** formed a *cartel*, or agreement, to set the prices at which they would sell crude oil and to divide up the world market. There followed a succession of large price increases, so that by the end of the decade, refined petroleum was selling for more than ten times what it sold for in 1970. These increases profoundly affected all economies, but particularly those of Western Europe, Japan, and the many less-developed countries that produce no oil of their own. They also sent waves crashing through the American economy.

Many factors contributed to the rise in petroleum prices during the 1970s. The increased price of OPEC crude was only one, but it was a major factor, and in many ways the most interesting aspect of the whole period, since it involved a major *international* redistribution of income. The losers were the residents of countries that were net importers of oil. People paid for this redistribution by having their gasoline and fuel oil prices rise by more than their money wages. They also paid for it in more subtle ways. Refined petroleum is a major input into the transportation industry, accounting for about 3 percent of transport costs in 1972, and for far more now that its price has risen. It is also a major input into electricity generation. Furthermore, petroleum products are a large part of construction costs, both as fuel for construction machinery and as the gooey part of asphalt. Altogether, there are 20 industries for which the cost of refined petroleum inputs was more than 1 percent of the price in 1972. Today, all their outputs are much more expensive than they would have been without the OPEC cartel, and their buyers are paying the bill.

This rise in the price of goods produced from crude oil need not by itself have caused an inflation, a *general* increase in prices. But when some prices go up, others must fall if overall prices are to remain stable. This would have required more downward price flexibility than has been characteristic of the American economy since the Great Depression. Another sizable depression might have been required for overall prices to remain stable. The government made an implicit policy decision not to risk such a depression in the name of price stability. It chose instead to let the *general price level* and the *relative price of oil* both rise. The alternative would have been to drive other prices (and money wages) down, achieving a relative price change without a general price change.

Before leaving OPEC, it might help you put events in perspective to think about the international redistribution from the point of view of the people at the other end. Most of the residents of OPEC countries are much poorer than the people you know. American popular magazines like to represent OPEC as a leering Arab wearing a mask and staging a stickup with the nozzle of a gas pump. Besides being racist, this caricature misses part of what

has been going on. The OPEC nations are using their oil revenues to develop their economies. For every oil sheik who plans to buy the New York Yankees for his favorite son to play with, there are dozens of planners, both capitalist and socialist, who are financing harbors, steel mills, and irrigation projects, and sending their young people abroad for technical and managerial training. They see how rich we have grown with the help of cheap energy pumped from beneath the surface of their lands, and they want to catch up to us. So far, relatively little of this new wealth has "trickled down" to the ordinary people of the OPEC countries, but they will be among the long-run beneficiaries if the cartel holds together.

Direct and indirect labor costs

Intermediate goods play a key role in *transmitting* price increases, but no discussion of input prices would be complete without an explicit discussion of money wages and labor costs. If you look back at Table 1, you will see that employee compensation seems to have been only about 27 percent of the cost of producing appliances in 1972. The dominant element was the cost of intermediate goods, which was about 58 percent of the price of the industry's output. But suppose that the 27 percent and 58 percent figures for the appliance industry were roughly representative of the industries supplying inputs to the appliance industry. Then 27 percent of the 58 percent would also have been labor cost in the supplying industries, and 58 percent of the 58 percent would have been intermediate goods used to produce the intermediate goods. These would also have had labor and intermediate goods as inputs to their production processes. All told, the direct and indirect labor costs of producing a dollar's worth of appliances would have been given by:

$$
\begin{aligned}
&+.27 && = \text{direct labor cost}\\
&+.27(.58) && = \text{first stage of indirect labor cost}\\
&+.27(.58)^2 && = \text{second stage of indirect labor cost}\\
&+.27(.58)^3 && = \text{third stage of indirect labor cost}\\
&+ \text{ etc.} &&\\
&= .64 && = \text{total labor cost.}
\end{aligned}
$$

The total labor cost of producing appliances would have been 64 cents on the dollar, about 2.4 times as large as the direct labor cost, because of the large amount of indirect labor required to produce the inputs to the appliance industry. The 64 percent figure is quite close to the value for industry as a whole. In 1972, employee compensation was 66 percent of the value of output produced by U.S. corporations. Labor costs vary between 65 and 70 percent of the value of output and are the dominant cost of producing GNP. ***The analysis of cost-push inflation is largely the study of changes in money wages.***

The importance of labor unions

Because labor costs are so important in determining prices, "Big Labor" is the villain in many an account of the cause of inflation. According to this scenario, monopolistic unions relentlessly drive up wages, squeezing the profits of their employers and forcing them to raise prices. Even though their greed is obviously self-defeating, it is boundless, and the rest of us can only look on as helpless victims.

This is an obvious exaggeration, but many people believe that it contains a big element of truth. Maybe they are right. But before you make up your mind once and for all, think about some statistics that will put union labor into quantitative perspective.

In this country, fewer than one out of four wage and salary workers belongs to a union. Union membership is heaviest in mining, construction, manufacturing, transportation, communications, and public util-

Holdup at the Gas Pumps: Was OPEC Guilty?

Between 1972 and 1980, the retail price of a gallon of regular gas went from about a quarter to about a dollar and a quarter. Most news media blamed OPEC, whose crude oil prices rose by a factor of 9.7 over the same period, from $3.50 a barrel to $34 a barrel. How just were the media in blaming OPEC?

The total U.S. supply of refined petroleum in 1972 was $33.6 billion. The attached diagram shows where this came from and where it went. Of the total refined supply, about 47 percent went to final demand; the rest went back into the input-output system. About 9 percent of this refined petroleum was imported. The other 91 percent was produced by the domestic refining industry. The value of domestic production was due partly to the crude oil that went into it, partly to other intermediate goods, and partly to value added. The crude oil came from both imports and domestic wells.

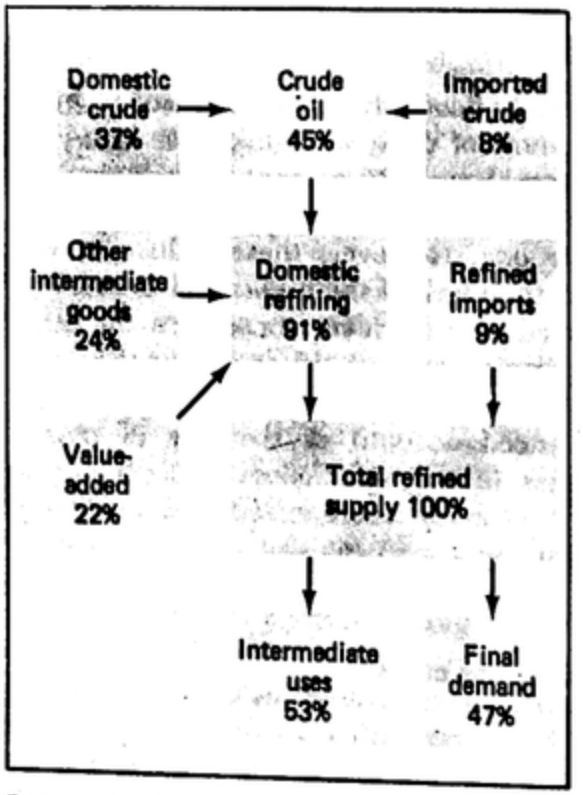
The percentage distribution of the costs and profits of domestic production and imports looked like this:

Domestic supply		91%
domestic crude used	37	
imported crude used	8	
other intermediate goods used	24	
value added	22	
Imported supply		9%
Total supply of refined petroleum		100%

These percentages can be used to compute a variety of price indexes that will help you understand why the price of gasoline went up as fast as it did.

Suppose, for instance, that the prices of imports of crude and finished petroleum both had gone up by a factor of 9.7, but that the prices of everything else affecting petroleum had stayed unchanged (that is, "increased" by a factor of 1.0). What would have happened to the price of finished petroleum? To get the answer, multiply 9.7 by 17 percent, the combined (8 + 9) weight of imports of crude and refined petroleum, and 1.0 by 83 percent, the combined (37 + 24 + 22) weight of everything else. When these two parts are added up, they come to 250. So prices would have risen by a factor of 2.5 if imports had been the only source of the inflation.

Suppose that besides the 9.7-fold increase in the price of imports, there had



Sources and uses of the value of refined petroleum, 1972

also been an increase in the price of domestic crude by a factor of 6.9. (This, in fact, was what happened when the world price went up.) Then the 1980 price index for refined petroleum (on a base of 1972 = 100) would have been:

$$\underset{\text{imports}}{17} \times 9.7 + \underset{\text{domestic crude}}{37} \times 6.9 + \underset{\text{refining costs and profit}}{46} \times 1.0 = 466$$

Thus, refined petroleum prices would have risen by a factor of 4.66, nearly as much as they did rise.

Now, suppose imports and domestic crude had risen just as in the previous example, other intermediate inputs into the refining process had risen in price by a factor of 2.3 (the PPI increase in intermediate inputs for manufacturing), and value added per barrel had risen by a factor of 1.8 (the actual increase in unit labor costs in private industry). Then the price index would have been:

$$\underset{\text{imports}}{17} \times 9.7 + \underset{\text{domestic crude}}{37} \times 6.9 +$$

$$\underset{\substack{\text{other intermediate goods} \\ \textit{(refining costs and profit)}}}{24} \times 2.3 \quad + \underset{\text{value added}}{22} \times 1.8 = 515$$

That is, prices of refined petroleum would have risen by a factor of about 5. That, of course, is just what happened.

Who, then, were the guilty parties? Imports accounted for about half of the price increase, domestic crude for most of the rest, and refining costs and profit for very little.

ities. Collectively, these industries make up most of what we think of as heavy industry. Nearly 40 percent of their workers belong to unions—a minority, but a sizable one. However, these industries pay less than half of the wages and salaries received by American workers. In the other industries, only 10–15 percent of the workers belong to unions. Clearly, then, any balanced account of the role of money wages in the inflationary process has to look at both unorganized and organized labor.

The Phillips curve

Most systematic analysis of wages and inflation starts with a discussion of the ***Phillips curve*** phenomenon. It is named for the British economist A. W. Phillips, who studied wages and unemployment in late 19th- and early 20th-century Britain. Figure 1 shows what his research uncovered. He found that if he graphed the yearly percentage change in money wage rates against the unemployment rate for the same year, his data described a curved, *inverse relationship*. Low unemployment rates went with high rates of money wage change. High unemployment rates went with low or even negative rates of change in wages. Such a curve described quite closely data for a long period of British history during which labor unions were insignificant in determining wages.

A similar curve does a very good job of describing the U.S. experience during the 1950s and 1960s, when prices as a whole were fairly stable. Figure 2 shows the relationship between changes in hourly earnings in the private economy and the unemployment rate for experienced wage and salary workers, the mainstream labor force referred to in the last chapter. Al-

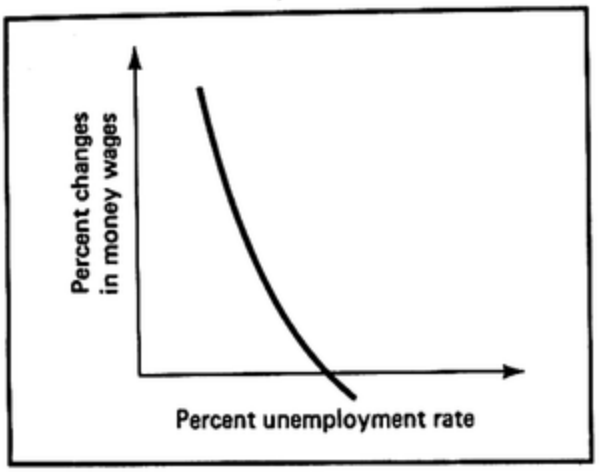


Figure 1 The Phillips curve

When A. W. Phillips studied the relationship between the unemployment rate and the rate of change of money wages in late 19th- and early 20th-century Britain, he uncovered an inverse relationship between the two. At low unemployment rates, wages tended to rise rapidly. At high unemployment rates, they remained steady or even fell.

though the data lack the precision of a laboratory experiment, you will agree that there is a relationship there that is not just the result of one or two outlying points. The slope of the straight-line approximation drawn through these points is −.75.

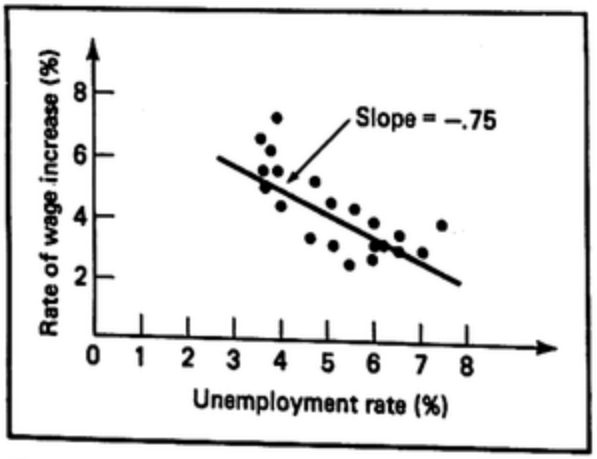


Figure 2 The Phillips relationship for the U.S. economy 1950–1969

During the 1950s and 1960s, the performance of money wages in the United States tended to confirm Phillips' relationship. On average, a 1 percentage point increase in unemployment was associated with a ¾ point drop in the rate of money wage increase.

Source: U.S. Department of Commerce, *Economic Report of the President.*

This means that every 1 percentage point reduction in the unemployment rate was associated with a ¾ point increase in the rate of growth of wages. Because of the evident curvature of the data, this figure applies only to employment rates in the middle of the graph, say between 4 and 6 percent. At lower unemployment rates, the effect of still further reductions appears even more inflationary. At higher unemployment rates, the effect of a reduction in unemployment seems quite small.

During the period described in Figure 2, economists and others in public life used to refer to the Phillips curve as a *trade-off.* It seemed to present a set of alternatives, ranging from rapid wage increase with low unemployment to fairly stable wages and a lot of unemployment. Arguments about social policy were couched in terms of moving toward a low unemployment rate at an "acceptable" rate of wage inflation, or toward more moderate wage growth at an "acceptable" rate of unemployment. The issue seemed to be the desirability of having more of one good thing at the expense of having less of another. This is a question of values—normative economics. The positive description of the range of possible choices was thought to be expressed in the Phillips curve.

The data on which Figure 2 is based end in 1969. Beginning in about 1970 and extending through the 1970s, the relationship between unemployment and wage increases changed dramatically. What had seemed almost like a natural law ceased to operate. Unemployment rose, and the rate of wage increase rose along with it. "Stagflation," a combination of stagnation and inflation, was upon us. Apparently either the trade-off had worsened for some reason, or the whole thing had been a historical coincidence all along.

There is good reason to think that in times of persistent inflation, relationships such as the Phillips curve governing wage

and price changes will shift upward. Persistent inflation gets built into people's expectations. When they expect prices to rise, they make different decisions than those they make when they expect prices to remain more or less stable. You will learn more about this later in the chapter. But unlike more recent years, the 1950s and early 1960s were not marked by persistent inflation, and the Phillips curve appears to have been fairly stable. Occasional spurts in the consumer price index (CPI) were separated by periods in which it was quite steady, hardly growing faster than the improvement in the quality of goods. The producer price index (PPI) was steady much of the time, as falling food prices offset rising industrial prices. With such a diversity in year-to-year price changes, it is doubtful if many people had strong convictions that prices in general would steadily rise, at least until the latter 1960s. Therefore, the trade-off between unemployment and wage increase was not obscured in the data by *changing* inflationary expectations. When expectations change, they shift the Phillips curve. If they change frequently, the curve shifts frequently. Historical data that come from a shifting Phillips curve do not describe a single trade-off, since each data point corresponds to a different state of expectations. But the trade-off is probably still there and would emerge in the data if people's expectations would stabilize around some consensus on the "normal" rate of price increase.

Why is there a Phillips curve?

Statistical regularity is not theory. Even if the Phillips curve exists, it must be explained. Since wages are determined in a mixture of unionized and nonunionized labor markets, the theory of the Phillips curve must have two parts, one for organized and one for unorganized labor.

In unorganized labor markets, the wage is regulated by supply and demand. Buyers of labor power maintain their work force by offering to pay a "competitive" wage, one just high enough to attract the necessary number of workers. If they can't keep workers or attract more when they want them, they offer to pay higher wages. When large numbers of people are out of work, firms can attract all the labor they want at the going wage. Workers are glad to have a job and have no reason to think they can get more pay somewhere else. When the pool of unemployed workers is small, firms that raise their wages to attract more workers find that this doesn't work. Other firms are bidding for the same inadequate supply. Workers can be choosy, searching for the best wage around. Disappointed firms then must raise their wage offers still further. The result of this process is that when unemployment is high, firms can achieve their goals without collectively raising wages. When unemployment is low, they are continually shorthanded even though they offer continually higher wages.

In organized labor markets, wages are determined by labor-management negotiations. The threat of a strike is always hanging over the negotiations. When times are bad, management is in a stronger bargaining position than labor. If there is a lot of unemployment, union members are already hurting financially, and strike funds are low. The additional cost of a strike makes a bad situation worse. Management, on the other hand, may find a strike less costly in bad times than in good, since a firm that is barely covering the costs of wages and materials has little to lose by shutting down altogether. Management can then stick to a low offer even if it provokes a strike. Labor may well decide to accept the offer rather than striking. In good times, the power position is turned around. A strike is very costly in terms of

lost profits, and labor is in a strong enough financial position to be willing to strike. The result is that management has an incentive to make a good offer to prevent a strike. Even a good offer may be refused in favor of a work stoppage that ends in a still more favorable settlement. Contributing to a higher settlement is the presumption that wage increases are easier to pass on to customers in good times than they are in bad.

Both in organized and in unorganized markets, then, the rate of wage increase is governed by the general degree of prosperity. When there is a recession with high unemployment, wages rise at a very moderate rate. If the demand for labor increases as a result of higher demand for goods, the unemployment rate falls, but the rate of wage increase rises only a little. The multiplier process can operate unhampered by the workings of the labor market. But when the economy gets very prosperous and the unemployment rate reaches a low enough level, particularly in the mainstream work force, wages start to rise rapidly. Higher demand for labor increases wages much more than it increases employment or output. The multiplier process, which works through changes in real output, gives way to the inflationary process, which works through changes in prices and money wages.

Productivity and the cost of output

So far, you have looked at the forces that bring about changes in the costs of inputs, including labor. However, input costs are only one element in determining output costs. The other element is how much output these inputs produce.

The arithmetic of costs

Most of the data on the productivity of inputs focuses on labor, simply because labor cost is so important to the economy. Think about the relationship among wages, labor productivity, and labor cost. Suppose that you work as a pin maker. Your boss pays you $10 an hour, and (using his machinery) you produce 1,000 pins an hour. How much is the labor cost of a single pin? Obviously, $.01—$10 per hour divided by 1,000 pins per hour. Your boss' ***unit labor cost*** (labor cost per unit of output) is one cent, equal to the ratio of your *employee compensation per hour* (your wage) divided by your *output per hour.*

The U.S. Bureau of Labor Statistics prepares data of this sort for the entire private economy, with and without the farm sector. Since farm prices are not very closely tied to costs in the short run, the data that exclude agriculture are the more interesting for studying inflation. The employee compensation data cover wages, salaries, fringe benefits, and employer contributions to social insurance, all divided by the total hours of labor employed. The output data measure constant-dollar GNP originating in the private nonfarm sector, again on a per hour basis. Unit labor costs are simply compensation per hour divided by output per hour. Changes in unit labor costs may be compared with changes in price, as measured by the GNP deflator for private nonfarm output. Changes in compensation and output per hour, unit labor costs, and prices over the 1950–1980 period are presented in Table 2.

As a matter of arithmetic, the percentage change in any ratio approximately equals the difference between the change in the numerator and the change in the denominator, at least for small changes. The first three columns of Table 2—compensation per hour, output per hour, and unit labor cost—present the breakdown of changes in unit labor costs from 1950 to 1980. By comparing the third column with the first and second, you can verify that the percentage change in unit labor cost

Table 2 Unit labor costs and prices, private nonfarm business, 1950–1980

	Average Annual Percentage Change in:			
Time Period	Compensation per Hour	Output per Hour	Unit Labor Cost	GNP Price Deflator
1950–1955	5.3	2.2	3.0	2.8
1955–1960	4.8	1.7	3.0	2.4
1960–1965	3.7	3.3	0.4	1.1
1965–1970	5.3	1.5	3.8	3.9
1970–1975	8.0	1.8	6.1	6.3
1975–1980	8.7	0.8	7.9	7.2

As a matter of arithmetic, the percentage change in unit labor cost approximately equals the difference between the percentage change in employee compensation per hour less the percentage change in output per hour. Because of the high relative importance of labor costs and the dependence between costs and prices, changes in price closely mirror changes in unit labor costs.

Source: U.S. Department of Commerce, *Economic Report of the President.*

approximately equals the change in hourly compensation minus the change in hourly product. From 1950 to 1955, the 3.0 percent increase in unit labor cost almost exactly equaled the 5.3 percent increase in hourly compensation minus the 2.2 percent increase in output per hour. You can also see that during the 1950s and early 1960s, about half of the rise in hourly employee compensation was offset by gains in output per hour. For the 1950–1965 period as a whole, employee compensation rose about 4.4 percent per year. About 2.4 percent of this was offset by productivity increases, so only 2 percent a year showed up as rising unit labor costs. During the late 1960s and especially during the 1970s, when compensation was growing very rapidly, unit labor costs shot up because productivity gains were too low to offset much of the rising input cost.

The apparent slowdown in productivity gains during the late 1960s and the 1970s has not yet been explained to anyone's satisfaction. To some extent, it was the result of a more rapid shift in the *composition* of output toward low-productivity service industries and away from high-productivity goods industries. To some extent, it also resulted from rising real costs of pollution control, product safety, and job safety regulations, which reduced the productivity of all inputs. But even after making generous allowances for these and other obvious factors, there remains a large unexplained drop in the rate of growth of output per hour. Whatever its cause, its consequences were costly.

Productivity, costs, and prices

The fourth column of Table 2—GNP price deflator—shows the price increases corresponding to the other data in the table. As you can see, they correspond very closely to the changes in unit labor cost. This relationship is partly the result of using five-year periods. Year-to-year changes in prices don't match up very well with yearly changes in unit labor costs. But the correspondence is nearly exact over the years. The 2 percent annual increase in unit labor costs from 1950 to 1965 was matched by a 2 percent annual inflation rate in the GNP deflator.

There is a strong temptation to use these data to argue that inflation comes from *cost push*—that the price level gets pushed up by *independently* rising labor costs. The pattern displayed in Table 2 is certainly consistent with this interpretation. But, in fact, a little analysis will show you that the data are also just about what you would expect to see if increases in the demand for output (demand pull) caused all price *and* cost increases. Imagine an unusually strong demand for goods. Firms would try to profit from this by hiring more workers and producing more goods. Thus, the higher demand for goods would cause a higher demand for labor. Unemployment would fall and wage rates would rise. What started as an unsatisfied demand for goods would end with both rising prices and rising labor costs. The cost rise would result from the demand increase in the goods market and would not be the initiating cause of the inflation.

It is important for you to see that rapid productivity growth offsets inflationary pressure, whether that pressure comes from the supply side of the input markets or the demand side of the output markets. The *offset to a cost push* is simple. If the unemployment rate is low, money wages will rise fairly rapidly, judging from the Phillips curve. If productivity is stagnant, the rise in money wages will be entirely reflected in unit labor cost. But if productivity is rising rapidly enough, the money wage increase will be offset, and unit labor costs will not rise. *The offset to a demand pull* is only a little more complicated. It occurs because rising productivity lets firms meet higher demand for their outputs without big increases in their demand for labor and resulting increases in wages. ***A slowdown in productivity growth without a slowdown in the growth of final demand puts more pressure on the supply of mainstream labor, bids up wages, and leads to demand-pull inflation in both costs and prices.***

The supply side: A wrap-up

This look at the cost side of the inflationary process has covered a lot of ground. Before turning directly to the demand side, let's summarize some of it in a series of general propositions:

1. Nearly all prices are directly influenced by costs of production.
2. The principal direct cost of production for nearly all goods and some services is the cost of intermediate goods. The input-output structure is therefore one of the most important conduits for transmitting inflation throughout the system of markets.
3. From the perspective of cost pressures, a rise in import prices has the same effect as a rise in the prices of intermediate goods of domestic origin. However, a rise in import prices imposes both a real cost and a cost in terms of inflation, since a rise in relative import prices diverts real income to foreigners.
4. Since intermediate goods are directly and indirectly the products of labor, labor costs are the major element in the cost of output as a whole.
5. Except during persistent inflation, there is a fairly regular inverse relationship between the unemployment rate and the rate of increase in money wages. This relationship, called the Phillips curve, implies that money wages rise rapidly with low unemployment and slowly with high unemployment.
6. Rising money wages can be offset wholly or in part by rising labor productivity. At high unemployment

rates, the offset is complete, so that unit labor costs are stable. At low unemployment rates, the rate of growth of money wages may be high enough to overwhelm the normal offset from productivity, and the cost of producing output rises. If productivity gains are below normal, this makes the cost increases even larger.

Inflation and the demand for goods

It might almost seem that a theory of inflation could be put together by looking at just the supply or input side, without mentioning final demand. It would go like this: Increases in input costs mainly originate either in the labor market or in the prices of imports. To the extent that these cost increases exceed productivity gains, they push up output prices. The input-output structure generalizes this cost push by transmitting it from industry to industry via the prices of intermediate goods.

There is nothing wrong with this story except that it is not very complete. First, inflation often *originates* from a shift in demand rather than from an input cost push. If input costs rise, they rise because of the demand change. They are not the initiating cause of inflation. Second, even when inflation involves a wage push, the low unemployment that triggers it is *derived from* a strong demand for goods. If demand is weak, the unemployment rate will be high, and there will be no cost push. Third, no discussion of inflation could be complete without some analysis of the effects of price increases on the demand for goods.

This section of the chapter focuses primarily on the relationship between the demand for goods and the demand for labor—the way in which demand pull produces cost-push inflation. It also explores the effects of inflation on the demand for goods.

Categories of unemployment

Economists usually think of unemployment in terms of three categories: ***frictional, structural,*** and ***cyclical. A person is said to be frictionally unemployed if he or she is between jobs, or has just entered the labor force but has not yet gotten a job.*** There are also frictionally vacant jobs. As long as people and jobs are mobile from place to place and firm to firm, this sort of unemployment is unavoidable. In fact, most economists consider this sort of resource mobility socially desirable.

Structural unemployment is a different matter. One reason for structural unemployment is that the skills of the unemployed don't match up well with the skill needs of growing industry, even though unemployed people and unfilled jobs exist in the same labor markets. But sometimes the industrial basis for employment in an entire region is wiped out. When the last of the big forests in Michigan's Upper Peninsula were logged off at the end of the 19th century, there were suddenly many unemployed loggers, saloon keepers, and railroad workers in that part of the country. It took years for the surplus population to move elsewhere, and for other industries to move in and employ those who remained. ***Meanwhile, there was a structural imbalance between the region's labor supply and the available jobs.***

Cyclical unemployment is the kind that is documented from month to month in the headlines. When GNP grows rapidly during the recovery from a recession, employment grows faster than the labor force. People who had been laid off or fired during the recession go back to work. There is a reduction in cyclical unemployment. When the next recession starts, there

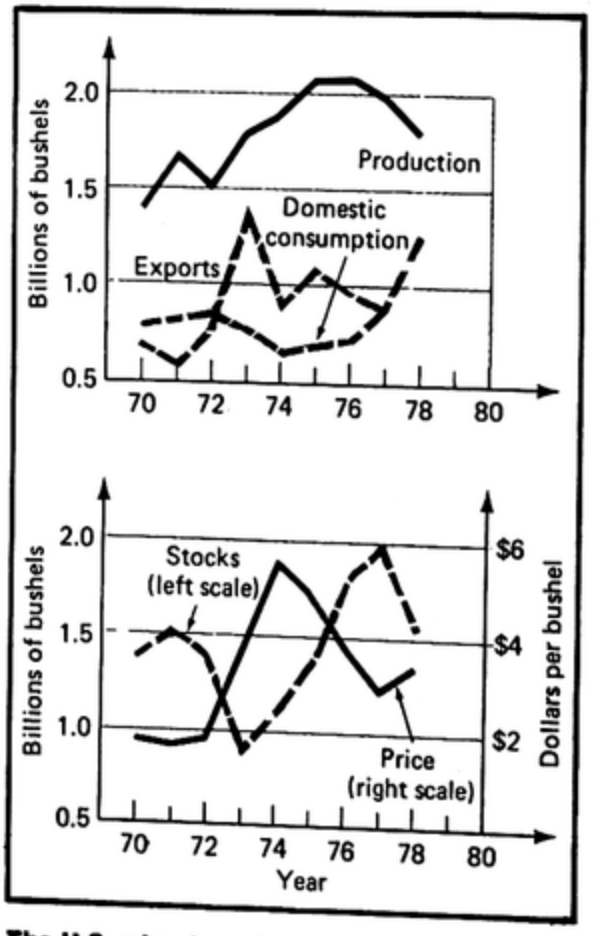
Demand Inflation: The Case of Wheat

History is full of inflations whose origins have nothing to do with cost increases. At the beginning of every war, people buy up goods in anticipation of shortages. What does this do? It creates shortages, and prices rise.

A good example of a demand shift with widespread effects occurred in the U.S. wheat market during the 1970s. The price of wheat increased by a factor of 3 between 1972 and 1974. So with the increase in oil prices that started at about the same time, our international trade was closely tied up with what happened. But instead of an import cost push, the rise in wheat prices was largely an export demand pull.

The United States is one of the world's major suppliers of wheat. In most years, we export about half of our crop. Among the other major suppliers are Canada, Australia, and the Soviet Union, which is also a major consumer. The weather causes the crops in the various producing countries to fluctuate sharply. The Soviet crop is particularly vulnerable to bad weather. Because wheat is such an essential commodity, the quantity demanded does not fall much when prices rise. Demand is *inelastic*. Thus, large price increases are needed to adjust the quantity demanded to a particularly small supply, or to contain a particularly large increase in demand itself. Year-to-year crop changes can be partly offset by changes in inventory stocks, but a couple of bad harvests in a row deplete stocks and send prices skyward.

During the early 1970s, the Soviet Union had several bad harvests. Since wheat is a staple of the Russian diet, the Soviet government had to go to the world market to supplement its limited domestic supplies. It negotiated a large deal with U.S. suppliers, shifting the demand for U.S. wheat. This deal was promoted by the U.S. government, which



The U.S. wheat market in the 1970s

The big wheat exports during the 1970s depleted stocks and sent prices up. Because wheat is such an important crop, it added to inflationary pressures during the years in which its price rose sharply.

Source: U.S. Department of Commerce, *Survey of Current Business*.

was trying to help the U.S. balance of payments by exchanging wheat for Russian gold. You can see what happened in the accompanying figure. Notice the sharp peak in U.S. exports in 1973, the year of the major shipments to Russia. The combination of these high exports and normal domestic consumption outstripped production and ran down stocks, which dropped by more than a third during the year. Prices rose sharply in 1973 and again in 1974. It was not until the large harvests of 1975–1977 that stocks rose and prices receded.

During this period, wheat rose from under $2 a bushel to nearly $6 a bushel. This was a response to the demand-supply balance in the world market for wheat, and not to an increase in the cost of producing grain. But, of course, it affected first the price of flour and then the price of bread. It thus affected the cost of living in a very noticeable way and definitely contributed to the development of the persistent inflation of the 1970s. Like the oil price increase, the wheat price increase *could* have been offset by falling prices elsewhere in the economy. But it was not.

is another round of firings and layoffs, and cyclical unemployment goes up. There does not even have to be a net drop in employment. As long as employment increases less rapidly than the labor force, cyclical unemployment will go up.

GNP, unemployment, wages, and prices

The most common form of demand-pull inflation involves both the labor and goods markets. Prices and money wages are pulled up together because final demand is larger than the economy's ability to produce.

For final demand to expand, production must expand. Suppose this happens. Then what? If production grows faster than the rising productivity of workers already on the job, employment will increase. What if this employment growth outstrips the growth in the labor force? Clearly cyclical unemployment will fall. This will bring the Phillips curve phenomenon into play. Money wages will grow faster. If the growth in money wages outpaces productivity gains, unit labor costs will rise, pushing prices up. What starts as an expansion in the demand for goods becomes a cost push from wages.

It may help you to understand this pattern of inflation if you know what would have to happen to let demand and output expand without rising prices. These conditions are illustrated in Figure 3. First, suppose that GNP in 1990 equals $3 trillion in 1984 dollars, represented by Point *A* in the diagram. Second, at Point *A*, the unemployment rate is 5 percent. Third, given the Phillips curve, this 5 percent unemployment rate is associated with a 2 percent rate of increase in money wages. Fourth, the rate of productivity gain is also 2 percent a year, so that unit labor costs remain constant at 5 percent unemployment. Fifth, the labor force grows at 2 percent a year.

Note that the first four of these conditions imply that at Point *A*, unit labor costs are steady. The 2 percent growth in money wages is offset by a 2 percent productivity gain. Suppose that the growth path from *A* to *B* represents a 4 percent expansion in final demand. This approximately balances the sum of 2 percent productivity and 2 percent labor force growth

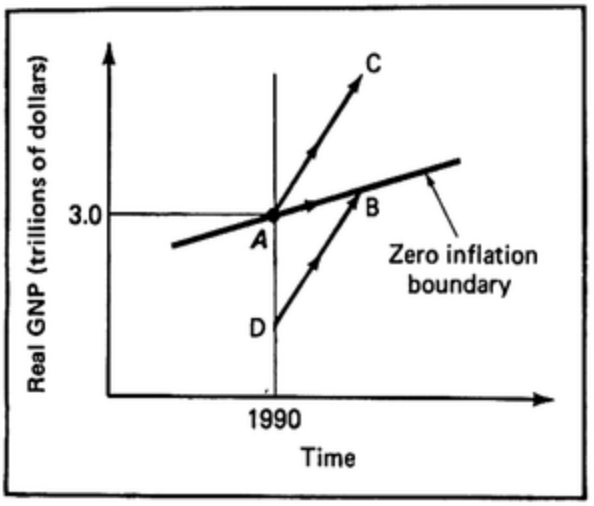


Figure 3 Sustainable and unsustainable expansion

If final demand and production expand along the zero inflation boundary, from *A* to *B*, the rate of growth in money wages is exactly balanced by productivity growth, and unit labor costs are constant. Prices are not pushed up. But expansion from *A* to *C* intrudes into the territory of the inflationary process. Expansion along the path from *D* to *B* would not be inflationary, but continued expansion past *B* at an unchanged rate would eventually lead to rising unit labor costs.

(ignoring compounding), so that when the economy reaches *B*, it will still have a 5 percent unemployment rate. It will therefore also still have steady unit labor costs. Starting from *A*, the economy could have a 4 percent expansion in GNP every year without any built-in upward pressure on prices. This is the kind of bedtime story that gives economists pleasant dreams. Everyone lives happily ever after along a time path like *AB*.

Suppose, however, that the economy expanded along path *AC* instead. Growth along *AC* would outrun labor force and productivity growth, and the unemployment rate would progressively fall. If Points *A* and *B* represent 5 percent unemployment and steady unit labor costs, then Point *C*, with its higher level of production, must represent a lower unemployment rate, money wage gains that outstrip the growth in productivity, and rising unit labor costs. The longer the economy proceeded along *AC*, the lower its unemployment rate would become, and the greater its rate of growth in money wages and labor costs would be.

Of course, if the economy were at Point *D* in 1990, its unemployment rate would be higher than 5 percent along a path such as *DB*, with falling unemployment but no inflation. However, once *B* was reached, its rate of growth would have to slow down, or it would enter the territory of the inflationary process. The expansion path *AB* is therefore in effect a *zero inflation boundary*.

Let's take a look at how this discussion stacks up against experience. Remember that this argument applies only when inflation is a sporadic, off-again-on-again thing, and not when there is persistent inflation. Think about the 1950s and 1960s, before inflation became a seemingly permanent part of American life.

Figure 4 illustrates how the GNP, employment, and inflation relationship worked out from 1950 through 1969. The top half of the diagram shows constant-dollar GNP relative to a boundary that is drawn through the data for 1950 and 1964, two years in which the unemployment rate was about midway between 5 and 6 percent. It slopes up at a rate of about 3½ percent a year—the combined rates of productivity and labor force growth over the period.

The bottom half of the diagram shows the overall unemployment rate and the year-to-year rate of increase in the GNP deflator, relative to the same horizontal axis used in the top half. Horizontal lines are drawn at an unemployment rate of 1½ percent. This corresponds to only 1 percent inflation in the private sector. Since this is not literally zero, think of the steady growth path in the top of the figure as a *low inflation boundary*.

To orient yourself, locate the 1950 and 1964 points on the top and bottom halves

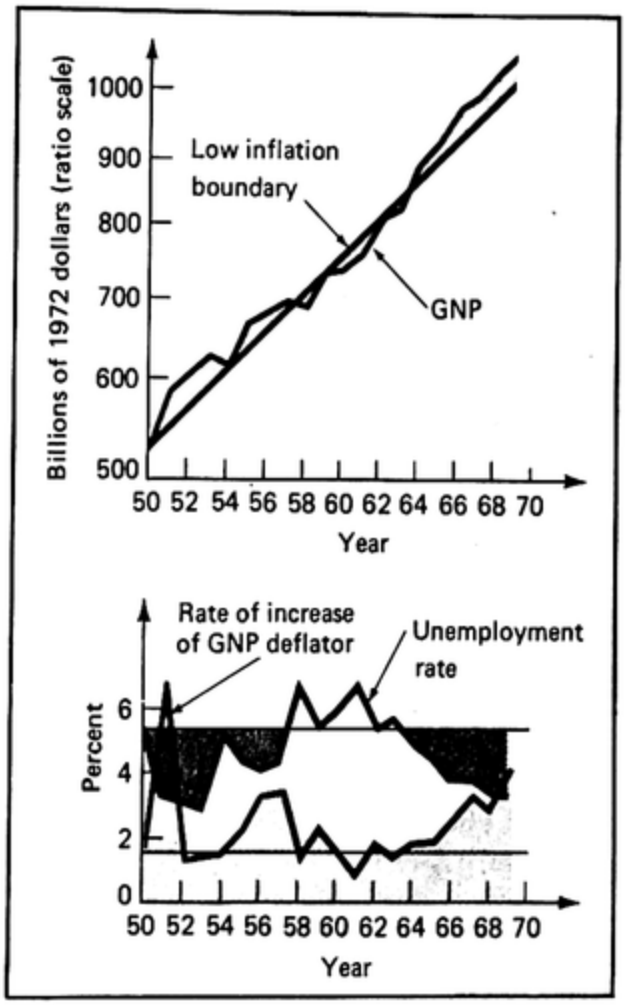


Figure 4 GNP, unemployment, and inflation in the 1950s and 1960s

During the 1950s and 1960s, the fluctuations in GNP were faithfully mirrored in changes in the unemployment rate. Since the Phillips curve relationship also worked, the rate of inflation in the GNP deflator went up and down as GNP rose and fell relative to the low inflation boundary.

Source: *Economic Report of the President.*

of the diagram. Verify that GNP for these two years was on the boundary, that the unemployment rate was about 5½ percent, and that the inflation rate in the GNP deflator was close to 1½ percent. Now study the rest of the diagram. There is an odd pattern in 1951–1953. This is the Korean War inflation, with a spurt of demand-pull inflation followed by two years of price controls. Aside from this unusual period, whenever GNP was above the boundary, the unemployment rate was below 5½ percent and the inflation rate was above 1½ percent a year.

Convincing isn't it? From 1954 onward, the relationship among the swings in GNP, unemployment, and inflation was about as regular as any complicated social process ever is. Because the rises and dips in the rate of increase of labor costs are so clearly associated with expansions and contractions in the goods market, it seems fair to think of the whole pattern as the usual way in which the business cycle pulls the price level with it at an alternating fast and slow rate of increase.

Prices and demand

Shifts in final demand, then, influence the general price level both indirectly through the labor market and directly in the goods market. As in any individual market, a rise in the price level feeds back on final demand and reduces it. But there is a very important difference between how the *general price level* affects final demand and how the *price of a single good*, say wheat, reduces quantity demanded in a single market. When wheat prices tripled in the 1970s, per capita wheat consumption dropped. It did so in part because people substituted other foods whose prices did not rise so fast for products containing wheat flour. Moreover, people whose money incomes grew more slowly than the price of wheat products (and this included almost everyone) suffered a real income loss and bought less of all goods, including wheat products. Wheat farmers and traders were gaining, but collectively they do not consume much wheat.

If you think for a while about GNP as a whole, you will see that the arguments that apply to a single good cannot be applied to all goods. There are no substitutes for everything, so that people cannot avoid high-priced goods by turning to something

else. Nor is there a direct loss in purchasing power. We know that GNY rises when GNP does. This is an accounting identity, and it must hold whether the GNP growth is mainly real output or mainly just prices. It is simply not true that inflation erodes the purchasing power of *everyone's* incomes. It results in redistribution, with some gainers and some losers. Therefore, it is a fallacy of composition to treat the inflationary process the way the newspapers often do, as though it were a real loss for all concerned.

People may, of course, see things the way the newspapers do. Most of us who get money wage increases think that they are well deserved, that they are the reward for doing a good job. We feel this way even if our raises really only reflect inflation. It is part of the charming and harmless psychic process of "rationalization." It makes us more comfortable with ourselves than we would be if we were forced always to see the truth. One of the nastiest aspects of inflation is that when our money wages are going up, rising prices take away the real gains that we think we have achieved through our merits. We see this as a triumph of the evil "system" over the worthy "individual"—ourselves. Of course, in inflationary times, rising money wages are just as much a part of the "system" as rising prices are.

There is, however, a wide range of economic quantities whose money values do not rise along with the price level. The most important of these is money itself. Whether in the form of currency or of bank deposits, money loses value when the price level rises. This relationship between prices and money is crucial to understanding inflation, and much of the rest of the book focuses on it. But until you take a close look at the banking system and where money comes from, it is not possible to tie up this particular loose end.

Persistent inflation

If you want to see how far wrong you can be if you think that one decade is much like another, look at Figure 5, and compare it to Figure 4. It shows the events of the 1970s relative to the boundaries used to analyze the 1950s and 1960s. What a difference! The unemployment rate more or less faithfully mirrors the relationship of GNP to the "low inflation boundary," but the Phillips curve seems to have disappeared. There is little if any association between unemployment and the rate of inflation in the 1970s data. And the average inflation rate is far higher than anything in the earlier decades, even though unemployment was consistently high.

That the relationships could change so much in such a short time shows that inflation is much more complicated than it looks in Figure 4. You need to look at several special factors to understand what made the 1970s different from earlier decades. By seeing how these factors affected the price level, you will gain some feeling for the kinds of complications that make Phillips curve theory incomplete, though not wrong as far as it goes.

What happened in the 1970s?

First, in the latter 1960s, unemployment was generally low and prices rose fairly fast, as you can see from Figure 5. This was a perfectly standard Phillips-type inflation. Demand for goods was very strong, led by government expenditures on the Vietnam War. This resulted in strong demand for labor relative to the labor force and sharply rising unit labor costs. By the beginning of the 1970s, prices had been rising sharply for several years. Then came the wheat and oil episodes which were piled on top of the Vietnam inflation. They show up as a two-year peak in the bottom half of Figure 5, at a time when high and rising unemployment rates might other-

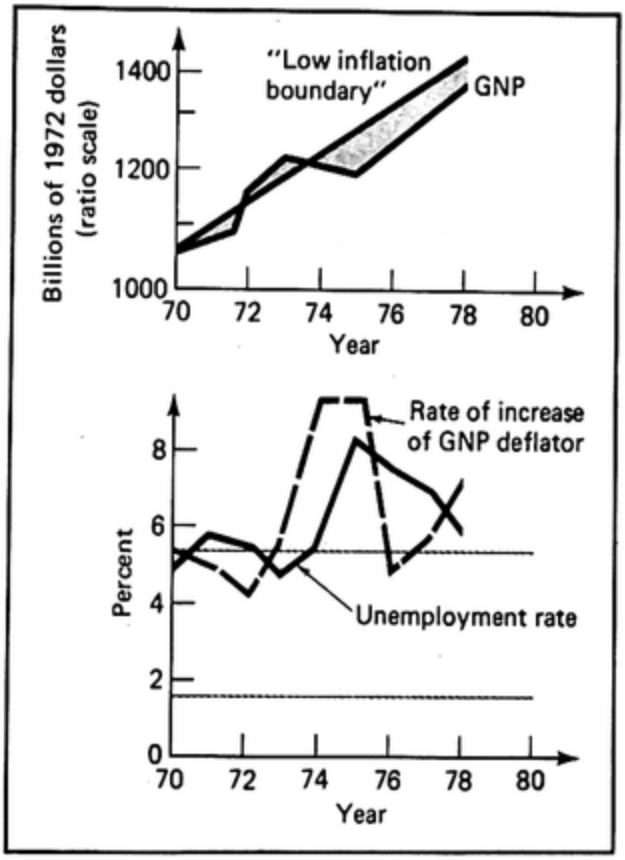


Figure 5 GNP, unemployment, and inflation in the 1970s

During the 1970s, changes in the unemployment rate continued to mirror changes in GNP, much as they had in the 1950s and 1960s. But the Phillips curve relationship broke down completely, so that the rate of change in the GNP deflator no longer bore any relationship to the unemployment rate.

Source: *Economic Report of the President.*

wise have lowered the rate of money wage increase a lot. From 1972 to 1975, the CPI for food increased 42 percent; for gasoline, motor oil, and antifreeze 56 percent; for home heating fuel 116 percent; for gas and electricity 41 percent. By 1975, real hourly earnings in industry, adjusted for changes in the CPI, were 3½ percent lower than they had been in 1972. Under normal conditions, they would have been about 5 percent higher, due to productivity growth. No one could fail to notice this, even if the news media hadn't given it such attention. Much of the loss in real wages has to be attributed to the food and fuel situations. Part of the income was redistributed to farmers. Part went overseas. The *causes* of these price increases had little to do with rising labor costs. But their *effects* influenced wage costs enormously over the rest of the decade. You have already studied how wage increases lead to price increases. Now think about how rising prices lead to rising wages.

Prices, wages, and expectations

Increases in the price level can influence wage rates in three ways:

1. Many wage agreements in unionized industries contain "cost-of-living adjustment" (COLA) clauses, under which money wage rates go up automatically to offset at least part of the rise in the consumer price index.

2. Losses in real wages increase the pressures on union leaders to bargain aggressively and may make the membership more willing to fight for a "living wage." Something similar happens in unorganized markets. Workers whose real wages are falling seek better jobs more actively. They mistake a general misfortune for something peculiar to themselves and think they can do better by changing jobs. This more intense job searching and the greater willingness of workers to quit forces employers to raise wages if they want to keep their work force.

3. In a period of prolonged inflation, people eventually accept steady price rises as normal. Both employers and employees then take continuing inflation for granted in all wage decisions. ***The decisions that they jointly make result in the very inflation they expect.***

This third point needs to be amplified. Suppose that you have a job that pays $5 an hour. The prices of what you buy and of what your employer sells have been steady for a long time, and both of you expect

them to remain steady. Your work is satisfactory and the job looks about as good as any you can get. The boss thinks you are a good worker, and wants to keep you. Wages in your area and industry generally rise about 2 percent a year. You and the boss both know this. To keep you, the boss must raise *your* wages 2 percent a year. A 2 percent raise is normal, and satisfactory both to you and the boss.

Now, suppose that prices, instead of being steady, are rising at 10 percent a year, both for what you buy and for what your employer sells. Instead of going up by 2 percent a year, wages in your area and industry go up 12 percent a year. You and the employer know all this. You still like the job; the boss still likes your work. What kind of a wage increase will he or she offer you? Of course, 12 percent. Whatever made the boss offer 2 percent at stable prices will lead to 12 percent at a 10 percent inflation rate.

You can see from this story how much expectations determine what happens in the labor market. A similar story could be told about how price agreements covering intermediate or final goods are also negotiated. What people decide to do depends heavily on what they expect to happen. And what, in fact, does happen is largely determined by what people decide. ***There is a big element of self-fulfilling prophecy in the inflationary process. If people expect inflation, they will do things that contribute to inflation. If they expect price stability, they will do things that contribute to price stability.***

The words "contribute to" are critical here. Wages and prices depend on much more than just expectations. For a given set of expectations, the more slack there is in the economy, the lower the rate of inflation will be. But it is perfectly rational for people to act on their expectations. When they expect prices to rise, they watch out for their own interests in ways that contribute to inflation. What seems advantageous for the individual will happen in the aggregate. It is only in the aggregate that it turns out to be self-defeating.

Expectational inflation is the kind of thing that makes social science so interesting and so different from the study of individual behavior. People get caught up in complicated social institutions that have rules of their own, different from the rules by which people regulate their individual lives. Adam Smith wrote about the many ways in which people's individual attempts to get rich produce benefits for everyone. But there are also many ways in which individuals and businesses cause public disasters just by trying to protect themselves against these very disasters. Think what happens when hundreds of people try to flee from a nightclub they think is on fire. The fire doesn't even have to be real to kill many of them.

It is hard to quantify expectations, since they aren't usually expressed in any direct form of behavior. They have to be deduced from indirect evidence. In looking at the persistent inflation of the 1970s, one indirect piece of evidence about changing expectations is the failure of the Phillips curve, after many years of working well. This *can* be explained by arguing that the mounting inflation got ingrained in everyone's view of the future. One result was to raise the rate of wage increase at any given unemployment rate, for reasons that we have just discussed. Another piece of evidence is the increasing willingness of public officials to admit that the inflation was going on and was likely to continue.

But the most convincing evidence of all is that the inflation was right there for everyone to see—if not in the news, then in the supermarket. Every time you went shopping, it hit you in the face. This was not true in the 1950s and 1960s. Inflation in those decades was intermittent, not persistent, year after year. In those days, if

you had predicted a high rate of inflation, you would have been wrong as often as not. In the 1970s, you would always have been right.

The shifting Phillips curve

Some economists argue that the Phillips relationship shifts upward by an amount that matches changes in the expected rate of inflation. The argument, which is illustrated in Figure 6, goes something like this.

Suppose that the Phillips curve *WW* corresponds to a situation in which people expect prices to remain stable. Each unemployment rate is associated with a certain rate of wage increase along with *WW*. If, for example, the unemployment rate is U^*, the rate of wage increase will be at a level indicated by W_A. Notice that at W_A, the rate of wage increase is just matched by the rate of productivity increase. Therefore, at U^*, there is no cost push pressure on prices.

Now suppose that demand for output increases, causing the unemployment rate to fall to U_L. At this lower unemployment rate, the rate of wage increase, represented by W_B, is greater than the increase in productivity. This will cause demand-pull pressures on wages and, therefore, cost-push pressure on prices. If this change in the unemployment rate is only temporary and the economy soon moves back to U^*, the Phillips relation will remain stable. Temporary changes in inflation do not lead to changes in expectations about inflation.

But suppose that demand remains high, so that the unemployment rate stays at U_L. The higher rate of inflation associated with U_L will come to be expected. Now in the process of wage determination, people will ask for and get a wage increase equal to W_B, plus an additional factor to compensate for the expected inflation. This wage adjustment for expected inflation causes the Phillips curve to shift up. At U_L, the rate of wage increase is no longer represented by W_B but by W_D. The influence of expectations, represented by the shifting of the Phillips curve from *WW* to *W'W'*, is added to the demand-pull inflation generated at U_L.

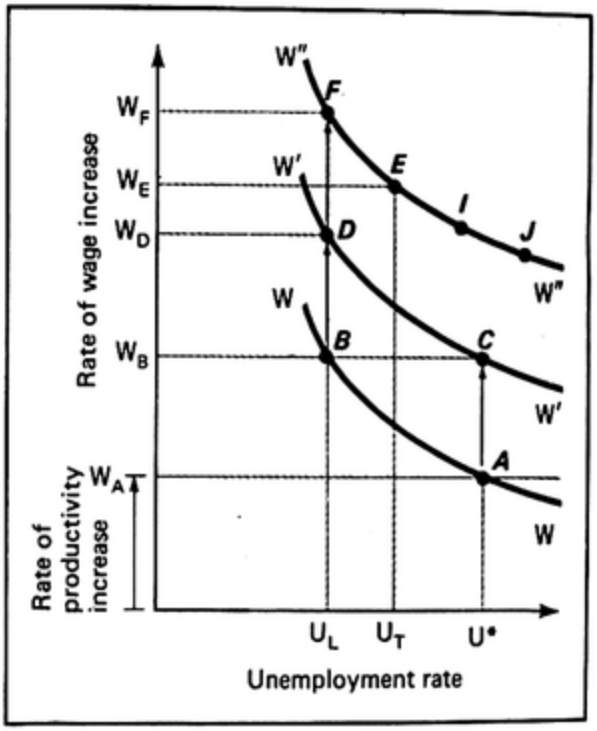


Figure 6 The shifting Phillips curve

Suppose that the Phillips curve *WW* corresponds to a situation in which people expect prices to remain stable. If the economy is located at Point *A*, with an unemployment rate of U^*, the expectations will be self-fulfilling. Suppose, however, that the unemployment rate falls to U_L, and remains there for a prolonged period. At Point *B*, the rate of inflation will be positive, since wages at *B* rise faster than productivity. Eventually, inflation will come to be expected, and the Phillips curve will shift upward by an amount equal to the shift in the expected rate of inflation. Now the rate of inflation at U^* is positive, not zero, because of the expectations of inflation.

Notice that the economy cannot remain at Point *D*, since the rate of inflation at *D* is higher than the expected rate that underlies *W'W'*. It can remain at *C*, however, since the actual rate of inflation at *C* just equals the expected rate underlying *W'W'*. In fact, the economy can have *any steady* rate of inflation at U^*, and *no steady* rate of inflation at a different unemployment rate. The rate U^* is the only one at which expectations are fulfilled. At any lower unemployment rate, the actual rate of inflation exceeds what is expected. At any higher rate of unemployment, the actual inflation rate falls short of what is expected.

With actual inflation higher, people's expectations about inflation will rise once more. If the unemployment rate stays at U_L, in the next round of wage bargaining, people will ask for W_D plus an additional factor to compensate for the higher expected rate of inflation. The Phillips curve will shift again. A still higher rate of wage increase will be associated with U_L.

At this point, policymakers may decide that something must be done about the increasing inflation. Knowing about the Phillips relation, they may decide to increase the unemployment rate, thereby lowering the rate of wage and price increase. To achieve lower demand and therefore higher unemployment, they may increase taxes. As the demand effects of the tax increase begin to show, the unemployment rate will indeed rise, perhaps to U_T. This will have a dampening effect on the rate of wage increase. However, this does not mean that the rate of wage increase will actually be lower.

Suppose that the unemployment wage increase combination is represented by Point D at the time of the tax increase. Even while the tax increase is causing unemployment to rise, people are responding to the recent increase in wages and prices by an increase in inflationary expectations. The Phillips curve will shift yet again, this time to $W''W''$. The unemployment rate–wage increase combination is represented by Point E. Without the tax increase, the unemployment rate would have remained at U_L, and the rate of wage increase would have been represented by W_F. But because of the rising unemployment, the rate of wage increase will only be W_E. The decrease in demand will thus have some impact. But if you compare Points D and E, you will see that E represents *both* higher unemployment *and* higher wage increases.

How can that be? ***The explanation is that the inflationary impact of rising expectations outweighs the deflationary impact of rising unemployment.*** Since inflation has continued to rise, the Phillips curve has continued to shift upward. Further increases in unemployment are needed. As unemployment increases, the rate of increase of wages and prices will begin to slow down. At some point, the deflationary impact of rising unemployment will be large enough to cancel out the inflationary impact of rising expectations. At that point, the actual rate of inflation will stop increasing. When the actual rate of inflation is stable, the expected rate of inflation will also stabilize. The Phillips curve will finally stop shifting. Further increases in unemployment will cause a movement down along the existing Phillips curve, as from I to J. The actual rate of wage and price increase will begin to drop. Once the actual rate of inflation begins to fall, the expected rate of inflation will also drop. The Phillips curve will start to shift down. Expectational inflation will begin to abate, but only at the cost of greatly increased unemployment.

A Phillips curve that shifts when expectations change alters the whole nature of the Phillips trade-off between inflation and unemployment. With a stable Phillips curve, there is a fixed trade-off. If unemployment drops to U_L, the rate of wage increase will rise from W_A to W_B along WW. As long as unemployment stays at U_L, the rate of wage increase will stay at B. But if the Phillips curve shifts, this is no longer the case. The trade-off for a lower unemployment rate is an ever-increasing rate of wage and price increase. The wage increase associated with U_L will be W_B, then W_D, then W_F. To reduce inflationary pressures, a simple movement back to U^* is no longer sufficient. A severe and prolonged recession may be necessary to generate sufficient deflationary pressures to counter the inflationary impact of rising expectations. The consequence of a persistent demand-pull inflation is a choice be-

tween ever-increasing inflation or a severe and prolonged period of high unemployment. This is a policymaker's nightmare.

The* theory of the shifting Phillips curve *makes it clear that actual inflation equals the sum of a demand-pull element and an expectational element. As long as actual inflation and expected inflation are not equal, then expectations about inflation, and therefore about the actual rate of inflation, must change. What unemployment rate, then, is compatible with a stable rate of inflation? If *actual inflation = demand pull + expectational inflation*, then actual and expected inflation can only be equal if demand pull is zero. Therefore, the only unemployment rate compatible with stable inflation rates must be that unemployment rate at which there are no inflationary pressures coming from wages. That must be the rate at which the rate of wage increase equals the rate of productivity growth. In Figure 6, that unemployment rate is U^*, where expected inflation and actual inflation are equal.

The economy can remain at C, for example, since the actual rate of inflation at C just equals the expected rate underlying $W'W'$. It can also remain at A, since the zero rate of price inflation at A underlies the curve WW. In fact, the economy can have *any steady* rate of inflation at U^*, and *no steady* rate of inflation at a different unemployment rate. The rate U^* is the only one at which expectations are fulfilled. ***Some economists call it the* natural rate of unemployment. *At any lower unemployment rate, the actual rate of inflation exceeds what is expected, and expectational inflation rises. At any higher rate of unemployment, the actual inflation rate falls short of what is expected, and expectational inflation declines.***

This argument, if correct, has important lessons for stabilization policy. It will be developed further in a later chapter. For now, just notice that it fits the experience of the 1970s fairly well. After a prolonged period of inflation starting with the Vietnam War, the rate of wage increase behaved differently from what had been typical in earlier years. With the exception of 1974 and 1975, the years of the worst food and fuel inflations, the GNP deflator moved in the opposite direction from the unemployment rate, but at an ever higher relative level.

Summary

About halfway through this long chapter, it seemed like a good idea to summarize the material on cost-push inflation. The first six of the points listed below just repeat what was said then. They all concern the role of costs in the inflationary process.

1. Nearly all prices are directly influenced by costs of production.
2. The principal direct cost of production for nearly all goods and for some services is the cost of intermediate goods. The input-output structure is therefore one of the most important conduits for transmitting inflation throughout the system of markets.
3. A rise in import prices has the same effect on cost pressures as a rise in the prices of intermediate goods of domestic origin. However, a rise in import prices imposes both a real cost and a cost in terms of inflation, since a rise in relative import prices diverts real income to foreigners.
4. Since intermediate goods are directly and indirectly the products of labor, labor costs are the major element in the cost of output as a whole.
5. Except during persistent inflation, there is a fairly regular inverse relationship between the unemployment rate and the rate of increase in money

wages. This relationship, called the Phillips curve, implies that money wages rise rapidly with low unemployment and slowly with high unemployment.

6. Rising money wages can be offset wholly or in part by rising labor productivity. At high unemployment rates, the offset is complete, so that unit labor costs are stable. At low unemployment rates, the rate of growth of money wages may be high enough to overwhelm the normal offset from productivity, and the cost of producing output rises. If productivity gains are below normal, this makes the cost increases even larger.

The remaining summary covers the second half of the chapter. It focuses mainly on the role of demand in the inflationary process.

7. Many episodes of inflation originate in particular markets and may be traced to sudden demand increases that cannot immediately be matched by supply. Such demand-led inflations may even be general, as in a wave of panic buying at the beginning of a war.
8. However, the most prevalent pattern of demand inflation intrinsically involves the labor market. Increases in the demand for goods lead to higher demand for labor, dwindling unemployment, rising wages, and increasing labor costs. Thus, higher demand for goods leads to higher costs.
9. Conceivably, final demand could grow at a rate that just equaled the sum of labor force and productivity growth. If this happened, the unemployment rate would remain constant. If it were constant at a level that maintained stable unit labor costs, then the demand growth could be achieved without rising costs and prices.
10. If final demand grows faster than the sum of labor force and productivity growth, it must lower unemployment, eventually producing a shortage of mainstream labor and pulling up labor costs.
11. The period of the 1950s and 1960s shows this process at work. The unemployment rate during these years closely followed the fluctuations in GNP. Periods of expansion produced low unemployment, money wage increases in excess of productivity gains, and rising prices. Whenever the expansion in GNP slowed down, unemployment rose, and inflation subsided.
12 During the 1970s, the pattern of events was quite different. The rate of inflation was exceptionally high by the standards of recent history, yet the unemployment rate was also very high. This change partly reflected the effects of special circumstances that produced exceptional increases in food and fuel prices. Even abnormally high unemployment could not have kept this from happening.
13. In part, however, it reflected the fact of persistent inflation. Whenever inflation is rapid and prolonged, people learn to expect it. When this happens, the Phillips curve trade-off shifts upward. The same unemployment rates that lead to stable prices when people don't expect inflation lead to rising prices when they do.

Key concepts

Persistent inflation
Cost-push inflation
Demand-pull inflation

OPEC
Phillips curve
Unit labor costs
Shifting Phillips curve
Expectational inflation
Natural unemployment rate

Questions for review

1. Look up the definition of *inflation* in your dictionary. After reading this chapter, would you accept or reject this definition? Explain.
2. Explain why *intermediate goods* are considered a major transmission route for spreading price increases throughout the economy.
3. A friend of yours who is also taking economics is puzzled. She points out that Table 1 in this chapter shows that 58 cents out of every dollar of output in the appliance industry goes for intermediate goods and services. Only 27 cents out of every dollar is spent on employee compensation. Yet she also notes that the chapter introduction mentions that labor is the major input to the production process. She feels that there is either a logical inconsistency or a misprint. Show your friend the way out of her dilemma.
4. The Phillips curve relation shows that the percent change in wages and unemployment rates are inversely related. Explain why this relationship might hold:
   a. for unorganized labor markets
   b. for unionized labor markets
5. The Phillips relationship for the U.S. economy (Figure 2) shows wages rising when unemployment is low (around 4 percent) and falling when unemployment rates are high (around 6 percent). True or false? Explain.
6. a. Analyses of the 1970s inflation pointed to low productivity gains as one important contributing factor to inflation. What is the connection between inflation and productivity gains?
   b. Would productivity gains be useful in offsetting both cost-push and demand-pull inflation? Explain.

· 27 ·

Financing the Circular Flow

As you read and study this chapter, you will learn:

- what the major functions of money are
- what fulfills those functions in our economy
- what goes on in financial markets
- what the major institutions of those markets are, and how they fit together with the rest of the economy

An occasional hazard of being an economist is having to be polite to the cocktail party acquaintance who asks, "You're an economist. What do you think is going to happen to the stock market?" For most economists, the only honest answer is "I don't know, and don't much care." Since that doesn't win any friends, most of us say something noncommittal but wise-sounding on these painful occasions. We also vow silently to stop going to cocktail parties dressed as economists.

The world of high finance sounds glamorous. It conjures up visions of beautiful people, daring moves and countermoves, millions risked on a thread, all played against a backdrop of sex, diamonds, champagne, and penthouses in Manhattan. This is how it is in the novels sold in supermarkets. For all that most of us are ever likely to know, this is how it really is.

You will be disappointed to learn, however, that this chapter is about low finance. The bread and butter of financial life is in mortgages, municipal bonds, inventory loans, and accounts re-

ceivable. The stock market is one of the financial markets, but its quantitative impact is suprisingly small. Financial history is occasionally written by great crises: Black Friday, the Bursting of the South Sea Bubble, and the Collapse of the Tulip Craze. But most of the time, it is written by clerks in ledger books, on punch cards, and on the magnetic disks of computerized financial institutions.

Low finance is about the flow of funds through ***financial markets.*** These funds circulate because some households, businesses, and governments find it advantageous to run surpluses. Others find it advantageous (or a regrettable necessity) to run deficits. You may recall from our earlier discussion that the actual surpluses must add up exactly to the sum of the actual deficits. But the surplus purchasing power doesn't move itself. This is accomplished by the financial markets. They complete the circular flow by moving funds from the surplus units to units that wish to run deficits.

Figure 1 will remind you of what kinds of funds are involved. It is based on national income and product figures from 1980. In that year, households saved $104 billion from their incomes, and state and local governments took in tax money that was $29 billion higher than their expenditures. These were the surplus sectors. This $133 billion moved through the financial markets to the deficit sectors—businesses that invested $62 billion more than retained earnings; the federal government, which had a $61 billion deficit; and foreigners, whose imports from this country exceeded by $10 billion the dollars they received from selling us their exports.

The magnitude of the flow of funds from surplus to deficit sectors gives some idea of the importance of financial markets. It is even more helpful to have a clear picture both of the nature of money and of the ***institutions*** that make up the financial markets. The first section of this chapter deals with the nature of money and with the definition and magnitude of the U.S. money supply. The second section examines both the various institutions that make up the financial markets and the types of financial instruments that facilitate the flow of funds.

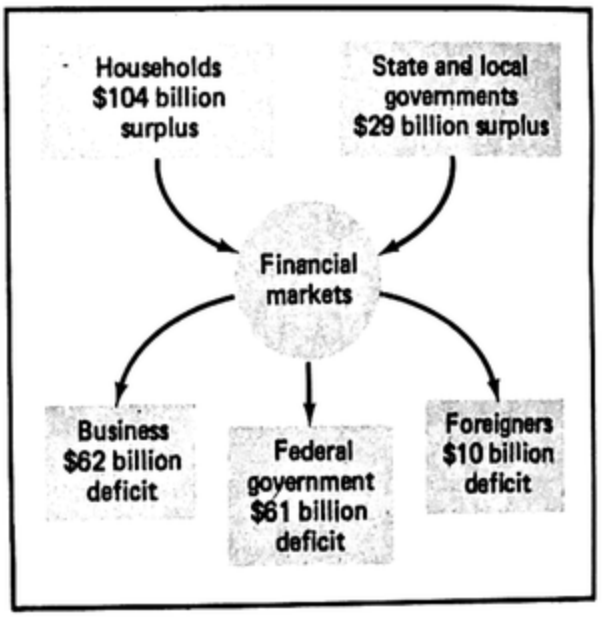


Figure 1 Completing the circular flow: Financing the deficits in 1980

In 1980, the business, foreign, and federal government sectors ran a combined deficit of $133 billion. This was financed through the financial markets. The funds came from households and state and local governments, whose combined surplus was also $133 billion.

Money

If all families were completely self-sufficient, trade or exchange would be unnecessary. ***Money*** as such would not exist, since it would not serve any economic function. As soon as there is barter or exchange, however, the development of some monetary unit is usually not far behind. Trading is much easier if there is some universally accepted token of exchange. Suppose you want to trade a sheep for some shingles. Think how much easier it is

to sell the sheep for money and use the money to buy shingles, than to find someone who wishes to trade shingles for sheep. Even the most primitive forms of exchange are usually carried on with some kind of money. In fact, the existence of a generally recognized symbol of power over goods may even preceed the development of specialization and exchange.

The functions of money

The most general way to define money is to list its functions. Physical form will not do, since a bewildering variety of objects have served as money throughout history: stones, shells, beads, metal, paper, cigarettes. But in a modern society, an object must fulfill several functions if it is to be considered full-fledged money.

1. Money serves as a universal ***medium of exchange.*** This is the most basic function of money. The presence of a token of exchange makes transactions easier in any economy, regardless of its level of development.

2. Because it is the medium of exchange, money serves as *the power to command marketable goods and services.* Firms, households, and governments borrow to obtain this power. In a market capitalist economy, money on the spot is the power to build a bridge, buy a house, take a trip, launch a ship or a new firm.

3. Money serves as a *universal equivalent or unit of account.* The monetary unit (the dollar, pound, mark, yen) makes it easy to compare the relative values of goods and services and to keep records. It is the measure of receipts and expenditures, gains and losses, assets and liabilities.

4. Money serves as a *store of wealth.* It is a very flexible form in which to hold wealth, since it is easily converted into other marketable forms, such as land, cattle, or corporate stocks. To serve well as a store of wealth, however, the value of the monetary unit in terms of goods must be fairly stable.

The uses of the *word* money are as varied as the functions of money. Look at the following statements. In each, money is used differently, yet correctly in at least one of its senses:

"You wouldn't believe how much money they take in every week." Here, money means the *universal equivalent,* a way of computing receipts. The same is true of the statement "John Lennon made a lot of money." Even the late Beatle couldn't "make" money, but he was paid a lot of it.

"I've got to write to my mutual fund and get some money." Here, the speaker wants to transform money as wealth into the form of a *medium of exchange,* so that it can be used to make purchases.

"Their money is all tied up in their business." In this statement, money is used as a measure of *wealth,* as a *unit of account.* "Tied up" money implies that wealth is not easily converted into money as a *medium of exchange.*

"Money isn't worth much anymore." This statement could refer to money as a store of wealth or a medium of exchange. In either case, it implies that the monetary unit won't purchase as many real goods and services as it once did.

"We'll have to go to the money market to finance this." Here, money means the *power to command goods and services.* The only purpose for borrowing in the money market is to gain this power to spend by obtaining some of the exchange medium.

These are some of the general ways in which the term *money* can be used. When economists talk about and study the *U.S. money supply,* however, they use a much more precise definition, based on the tangible objects that serve these functions in our society.

The U.S. money supply

One way to decide what makes up the supply of money in the United States is to ask what serves as a medium of exchange in American economic life. For example, what can you use to pay for something in a store? The only things that will work are currency (paper money and coins) and checks. Before you are tempted to add charge accounts and credit cards to this list, remember that they don't end the transaction. Eventually, you'll get a bill in the mail and will have to settle your account by writing a check. Credit cards and charge accounts are simply promises to pay. It is the sum of cash and checking accounts, all of which are means of exchange, that make up the most basic definition of money. Figure 2 shows the size and composition of two measures of the money supply used in the official reports of the Federal Reserve, the country's principal agency for regulating its monetary system.

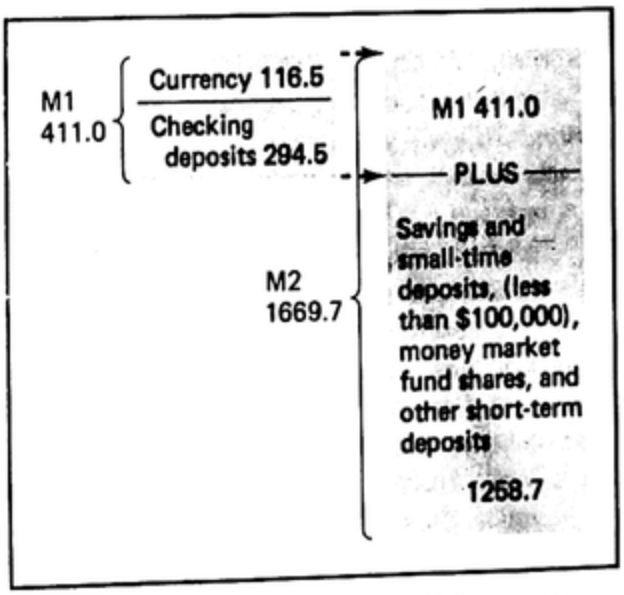


Figure 2 Two measures of the U.S. money supply December 1980 (in $ billions)

M1 and M2 are two of the most useful measures of the U.S. money supply. They are prepared by the Federal Reserve, the major controller of our monetary system.

M1 is made up of currency (coins and paper money) and all checking accounts in banks and savings institutions.

M2 equals M1 plus funds held in savings accounts, small time deposits, money market funds, and other various short-term deposits.

Source: *Economic Report of the President.*

The first block of Figure 2 is labeled M1. It is made up of currency and checking deposits, both in ordinary banks and in savings institutions. The M1 definition of money is most helpful in discussing the circular flow, since all transactions involve transferring M1 from one party to another. Most large transactions are made by check. In these cases, bank balances are simply transferred from the person or firm that writes the check to the person or firm for whom the check is written. Since the supply of bank money is continually passed around, many transactions can be financed by a small amount of money. If you compare the total value of checks written per year with the average amounts of money held in checking accounts, the checking account portion of the money supply seems to change hands over 100 times a year. In fact, deposits in the major New York banks, which handle the transactions of many large businesses, turn over more than 1000 times a year.

M1 is only one of several measures of the money supply regularly used by public officials, economists, and members of the financial community. Another important measures is M2. Besides the components of M1, M2 also includes a variety of savings and time deposits. ***Savings deposits*** pay interest steadily and may be withdrawn at any time. ***Time deposits*** mature at a definite date and entail an interest penalty if they are withdrawn earlier. Some of these, particularly certificates of deposit and money market fund shares, bear interest rates that are high enough to be competitive with those earned by other forms of wealth, such as corporate bonds. They are also particularly *liquid:* They can be quickly and easily converted into spendable money, sometimes by a phone call.

This makes them an attractive place to put short-term funds when interest rates are high. Since their owners consider them to be close substitutes for cash, any analysis of money and inflation must focus on M2, not just on currency and checking deposits, which make up M1. Figure 3 shows the trends in M1 and M2 from 1960 to 1980.

Money as debt

At one time, money was gold and silver coins, whose values as money were comparable to the costs of mining, refining, and minting them. Because such money had to be produced at some cost, it was both the medium of exchange for goods and a good itself. It used to be said that it had *intrinsic value.* Nowadays, money is nearly all made up of pieces of paper, or spots on a magnetic tape that keep track of checking account balances. It is *fiat money,* money by decree, without intrinsic value. What gives this money a value that is way above its cost of production?

If you sit down to read what it says on pieces of money, you will get some clues.

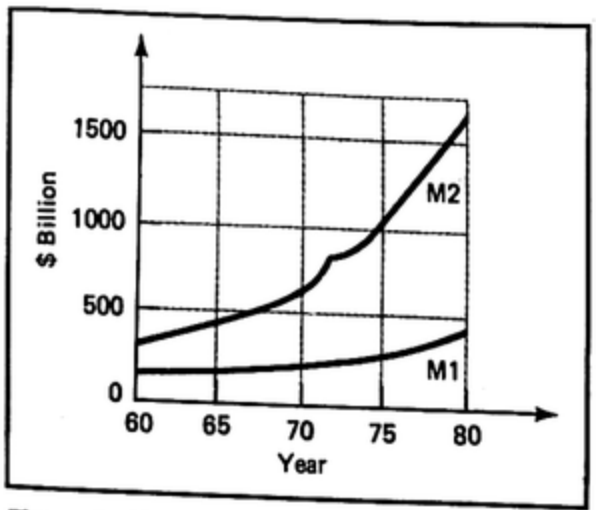


Figure 3 Measures of the U.S. money supply 1960–1980

An uptrend in the ratio of M2 to M1 in response to the high interest rates of the late 1960s and the 1970s shows up clearly in this graph.

Source: *Economic Report of the President.*

Take a check, for example: It is a command from the signer to a bank. It orders the bank to pay a certain sum to the "payee," the person or institution whose name follows the words "pay to the order of." The bank will carry out this order because it is in debt to the writer of the check. It "owes" the writer any money that the writer has on deposit with the bank. The technical name for a ***checking deposit*** held in an ordinary bank is a ***demand deposit:*** The debtor (i.e., the bank) must pay off the debt on demand, either to the depositor or to anyone else whom he or she designates by writing a check. Thus, the money represented by demand deposits is the *debt of a bank.* It is a liability to the bank but an asset to the depositor.

Currency makes fairly interesting reading. A dollar bill, for example, displays on its green side a lot of mystic symbolism, a profession of faith in God, some ceremonial Latin phrases, and extensive baroque ornamentation. On the gray side, it has a picture of Washington and the signatures of two high officials of the U.S. government. Dollar bills only circulate a few months on average, since they are destroyed when they become worn or dirty. The fame of these two officials is therefore fleeting. At the top, it says that it is a *Federal Reserve Note.* This identifies it as a *debt of the Federal Reserve System,* the branch of the U.S. government that manages the money supply.

Coins are much less interesting, at least to the casual reader. The Latin and the profession of faith are also there, along with the word "Liberty," but there are no clues to the source of the coins, other than their nationality. In fact, coins are issued by the U.S. Treasury. Nowadays, the value of the metal in most coins is not very great, except for that of pennies, which will probably not be made from copper much longer. During the past century or so, coinage has evolved from being intrin-

sically valuable to being just a noisy form of Treasury debt. All modern U.S. *coins are debts of the Treasury.*

Note that while you consider your money to be an asset, each of the three forms of M1 is someone else's debt. When you pay money to someone, you simply give that person title to the debt of a third party, either a bank (demand deposits), the Federal Reserve (paper money), or the Treasury (coins). And if the person whom you pay deposits this receipt in a checking account at a bank, he or she has decided to hold money in the form of a debt of that bank.

Modern money is just a symbol of indebtedness, therefore, but it can move mountains. Where does it get this power? Curiously, it gets it from people's belief that it is money. If people were to stop accepting debt as money, it would stop being money, right then and there. If everyone decided to accept only feathers as a means of payment, then feathers would immediately become money. Or if people were to choose some truly preposterous symbol—say, a soft yellow metal found only in a few places on earth and very expensive to mine, refine, and mint—then gold would be the king of the realm of transactions. But then, people would never be so foolish, would they?

Financial institutions and assets

All markets circulate money. In some, goods and services go one way, and money the other. In the financial markets, money goes one way, and promises to pay go the other. Some transactions just exchange one kind of money for another. If you write a check on your account at a bank and put it in a money market mutual fund, you are just restructuring your holdings of M2. But if you write a check to buy a U.S. Treasury Bill, you are exchanging money for a financial asset that is not money. There are many such assets sold on the financial markets.

Financial assets

The variety of paper assets is astonishing. They have one common characteristic, however. All promise to pay money in the future. Most financial assets fit into one of three categories:

1. ***Loans*** are promises to repay principal and interest according to a definite schedule. In legal terms the loan is called a "note."

2. ***Negotiable debts*** are also promises to repay on a definite schedule. But unlike notes, they may be sold by the initial purchaser to someone else—they are "negotiable." The buyer then gets the right to receive the payments promised thereafter. This category includes bonds, some mortgages, and a variety of other promises to pay. Most government securities held by the public fall into this category.

3. ***Equities*** are promises to pay to their buyers a share of the issuer's profits in the form of dividend payments, when and if such dividends are "declared." The best-known kinds of equities are corporate stocks. These are negotiable—that is, they can be resold.

Firms, households, and governments may finance deficits for a while by drawing down their money holdings. But if they run large or chronic deficits, they usually have to sell promises to pay in the future. Households are restricted to taking out loans, but both firms and governments can issue negotiable debts, and firms can sell equities. Foreign firms and governments can also borrow and sell securities in this country.

Firms, households, and governments that run surpluses can buy financial assets as well as accumulate money. The household sector characteristically runs a surplus. Some households channel their funds into savings accounts and time deposits. In effect, they allow financial institutions to invest for them. In doing so, they accept a low rate of interest, but avoid having to plan their investments carefully. Other households buy stocks and bonds directly or through a mutual fund. The business sector as a whole is usually in deficit, but individual surplus firms often buy short-term government securities to earn interest on their funds. The federal government is also usually in deficit, but state and local governments accumulate large surpluses in their employee pension funds. These are typically invested in negotiable debts, such as government bonds. The governments, firms, and individuals of other countries buy securities issued by U.S. firms and governments.

All the sectors of the economy, then, participate in financial markets, buying and selling financial assets. They also buy and sell goods and services. Their financial transactions, including the ups and downs in their bank accounts, just balance off their transactions in the markets for goods and services.

This process of balancing is much easier than it might be because of the existence of specialized financial institutions, many of which are familiar to you. The most familiar of these, and in many ways the most important, is the commercial bank.

Commercial banks

Commercial banks circulate the lifeblood of the business system. Their headquarters are usually in the downtown business districts of the cities and towns they serve. They handle all the checking accounts for business firms, and most of those that people use to carry on their transactions with the business sector. In total, about 80 percent of checking account deposits are held in commercial banks. Since these deposits turn over much more rapidly than checking deposits in other financial institutions, nearly 100 percent of the value of all checks is written on commercial bank accounts. At the end of 1980, about one third of the total deposits of commercial banks were checking deposits. Most of the remainder were interest-bearing time deposits, many of which were held as deposit certificates belonging to business firms.

Deposits are liabilities of the commercial banks. Their assets consist primarily of business and consumer loans, but also include security holdings. The predominance of loans reflects their overwhelming involvement in the financing of ordinary business transactions.

Commercial banks play a pivotal role in determining the size of the M1 money supply, since they are the major institutions that take checking deposits. You will learn a lot more about this in the next chapter. But there are several other kinds of financial institutions that *collectively* outweigh commercial banks in financing the deficit sectors of the circular flow. If you look at Table 1, you can see that only about 40 percent of institutional lending in 1980 came from commercial banks. The remaining 60 percent came from other insti-

Table 1 Funds loaned by private financial institutions, 1980 (in billions of dollars)

Commercial banks	103.5
Thrift institutions	57.6
Private insurance and pension funds	76.4
Other	28.1
Total	265.6

Source: *Federal Reserve Bulletin.*

tutions that are described in the following sections.

Savings institutions

Most small savers have very little contact with the security markets, and lack the funds, know-how, and courage to buy negotiable debts or equities. They like to save regularly and safely in small amounts and to be able to withdraw their savings on short notice. To earn interest, they put their savings in a savings or time deposit.

Both time and savings deposits can be kept in commercial banks, but several other institutions specialize in deposits of this kind: savings and loan associations, mutual savings banks, and credit unions. All accept small deposits, pay interest or dividends, and allow their depositors to withdraw money on fairly short notice without a lot of red tape and expense. Such ***savings or thrift institutions*** make consumer loans and deal in the securities and mortgage markets, accumulating diverse assets. In a sense, they invest money for the small savers. Indirectly, their depositors get some of the yields from such nonmoney assets as mortgages and government securities, without the risks and headaches that would be involved if they tried to buy them directly. Because the savings institutions diversify their assets, deposits in them are quite secure. Most of these deposits are also insured by the federal government, protecting them against anything but a catastrophic collapse of the whole economic order. Mutual funds perform a similar service for savers who have more to invest and are willing to take some risk to get a return higher than that available at the thrift institutions.

Table 1 shows that private insurance and pension funds are another major channel through which funds move from surplus to deficit sectors. Employees "contribute" to pension funds as a condition of employment. These funds are generally administered by life insurance companies. Withdrawals on demand are usually not permitted. Employees get access to their savings only when they retire. Just like thrift institutions, though, pension funds accept small deposits and use them to buy a variety of income-earning assets.

By now, you should see why commercial banks and savings institutions are known as ***financial intermediaries.*** They accept deposits, make loans, and buy negotiable securities. Without them, individual saver-investors would have to deal directly with individual borrowers and issuers of securities. Intermediaries allow savers to make a secure return on their funds without ever having to confront a borrower or judge if he or she is a good risk.

The security markets

There are other major financial institutions most savers rarely contact directly. A ***security exchange*** like the New York Stock Exchange is one example. Security exchanges are simply markets on which negotiable equities and debts (stocks and bonds) are bought and sold. Linked to the exchanges is a network of thousands of ***brokers and dealers*** with offices throughout the country, who aid in the buying and selling of securities. Brokers act as agents for others, carrying out orders to buy or sell. They get their incomes from brokerage fees. Dealers buy and sell on their own and make their incomes from the difference between buying and selling prices.

Buried deep inside the labyrinth of financial institutions is the *investment bank,* actually a specialized kind of security broker-dealer. These firms handle "new issues," that is, large blocks of new securities offered for sale by businesses and governments. Acting either as a middleman or the initial purchaser, the investment bank "places" the newly issued securities with other financial institutions that

want to buy them, or offers them for public sale.

Security dealers, brokers, and their customers are at the opposite end of the financial spectrum from the small saver who puts her or his funds in a financial intermediary. They have to know the risks and financial returns that are inherent in dealing directly with borrowers.

The flow of funds

Financial sectors and production sectors

Now that you are familiar with the major financial institutions, you can see exactly how they fit together with the production sectors. The links between the financial institutions and the production sectors are shown in Figure 4. In this example, the household sector is running a surplus, while the business, foreign, and government sectors are collectively running a deficit. The figure shows how the household sector finances the deficits of the other sectors.

If the household sector has a surplus, its receipts from production must be greater than its expenditures. This shows up as a net ***flow of funds*** from the other sectors to the household sector, along the bottom of Figure 4. Somehow these funds must get back to the business, foreign, and government sectors. Some of them return *directly*, through household purchases of government, foreign, and corporate securities. This is shown by the flow of funds from the households through the security markets, and on to the deficit sectors. The rest of the household surplus returns *indirectly*, through financial intermediaries. In this case, households deposit money in accounts with financial institutions, such as banks and mutual funds. The financial intermediaries use this money to make loans to businesses and to buy stocks and bonds, both domestic and foreign.

The U.S. flow of funds account

To see how important financial intermediaries are, look at Figure 5, which shows the movement of funds from deficit to surplus sectors in the United States during 1980. It is arranged a little differently from Figure 4 because grouping the economy into household, business, and government sec-

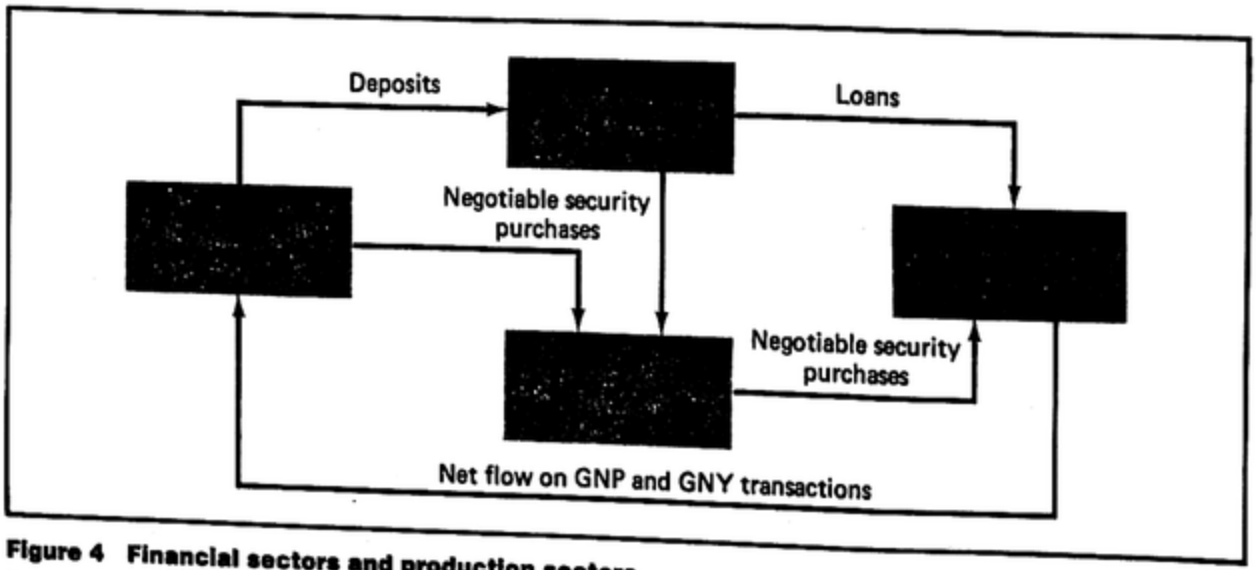


Figure 4 Financial sectors and production sectors
If the household sector runs a surplus and the other sectors run deficits, funds will be transferred to the deficit sector through financial intermediaries and the security markets.

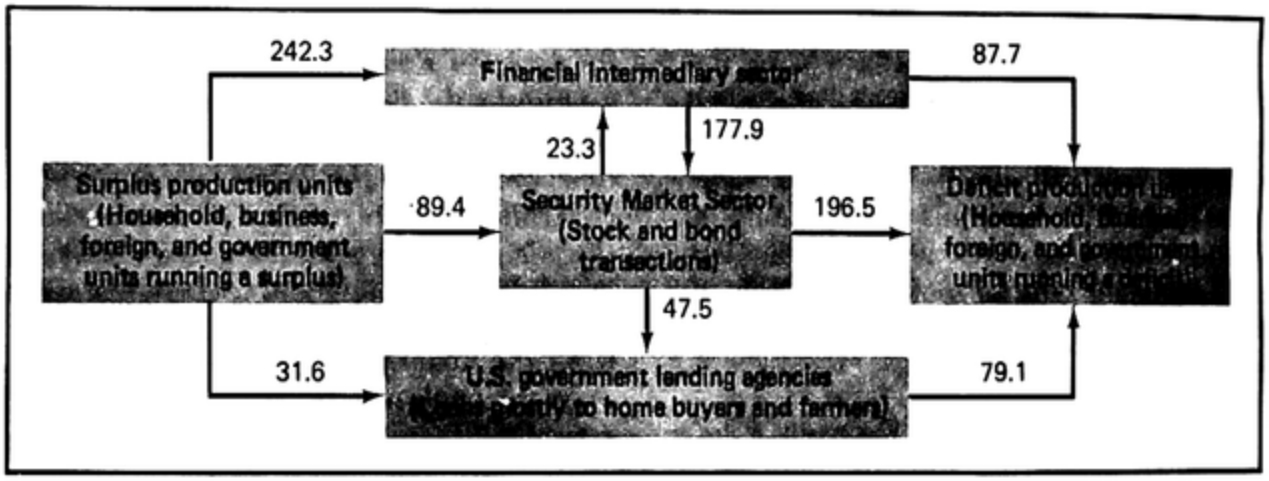


Figure 5 Flow of funds 1980 (in billions of current dollars)

In 1980, most funds that were transferred from surplus to deficit sectors went through financial intermediaries. Although deficit units sold $196.5 billion worth of negotiable securities, surplus units only bought $89.4 billion worth. The rest were acquired by intermediaries and financed by their deposits.

Source: *Federal Reserve Bulletin.*

tors understates the *total* volume of financial transactions. For example, showing a net surplus figure for households hides the fact that some households are saving, and some are borrowing. Total transactions, which include intrasector transactions, are larger than net transactions. Figure 5 divides the economy into all household, business, foreign, and government units running a surplus and all those running a deficit.

Part of the flow from surplus to deficit units in 1980 went through federal lending agencies that channeled taxes and their own borrowed funds into loans for farmers and home buyers. The rest went through private intermediaries and security markets.

In 1980, the total flow of funds from surplus to deficit units was $363.3 billion, nearly 15 percent of GNP. Most of these funds ($242 3 billion) reached deficit units *indirectly*, through financial intermediaries. Another $31.6 billion went to government lending agencies, so that only $89.4 billion went directly into the security markets. However, deficit production units raised $196.5 billion on these same markets. The remaining funds came from financial intermediaries, and indirectly from their depositors. In fact, the flow into the security market from intermediaries was large enough to enable other intermediaries and government lending agencies to raise $23.3 billion and $47.5 billion besides the funds raised by the deficit production sectors. Thus, while surplus units mainly build up deposits, deficit units take out loans and sell securities. It is the business of the financial intermediaries to convert the kinds of paper assets that some people want to sell into the kinds that others want to acquire.

Summary

This chapter has focused on the financial markets that link surplus and deficit sectors together. The major points to remember are the following:

1. Every economy with a division of labor and exchange has something that serves as money. This money functions

as the medium of exchange, gives people power over goods and services, provides the unit in which the accounts are kept, and serves as a store of value.

2. In the American economy, these functions are performed by the currency and checking deposits that make up what economists and financial analysts call M1, the medium of exchange. A broader definition of the money supply—M2—includes other kinds of deposits that can be quickly converted into equivalent amounts of M1.
3. The economy's surplus and deficit sectors are linked together by a network of financial markets that, in effect, complete the circular flow. On these markets, money is exchanged for promises to pay, such as loans, mortgages, stocks, bonds, and similar contracts for future payments. The suppliers of money (the surplus units) are those whose income is greater than their expenditures. The demanders of money (the deficit units) are those whose income is less than the amount they wish to spend.
4. Some of the lenders on the security markets deal directly with borrowers. But most funds flow through financial intermediaries. These are banks, savings institutions, pension funds, life insurance companies, and the like. Financial intermediaries accept the deposits of savers and use them to deal on security markets and to lend directly to deficit units. Depositors get safety and flexibility along with a return on their funds.

Key concepts

Financial markets
Institutions
Money
Medium of exchange
M1, M2
Checking or demand deposits
Savings deposits
Time deposits
Currency
Financial instruments
Loans, negotiable debts, equities
Commercial banks
Savings or thrift institutions
Security exchange
Brokers and dealers
Flow of funds

Questions for review

1. a. Which of the following items can be considered money, according to either the M1 or M2 definition? *Explain why* you would include or exclude the following items in a definition of money.
   i. a savings account
   ii. a share of GM stock
   iii. a checking account
   iv. coins
   v. a government security
   vi. a loan
   vii. a certificate of deposit
   viii. currency
   ix. a municipal bond
   b. Which are the most important components of the U.S. money supply in terms of volume? Consider both M1 and M2.
   c. What gives these components of the U.S. money supply their value?
2. What kind of financial asset (loan, negotiable debt, or equity) is involved in each of the following transactions?
   a. the purchase of a bond issued by AT&T

b. the resale of a mortgage
c. household borrowing to finance home remodeling
d. the purchase of IBM stock
e. the purchase of a government security
f business borrowing from Citibank to finance expansion of facilities
g. parents' borrowing to finance children's college education
h. a woman taking out a mortgage to finance the purchase of a home

3. The household sector is said to be a surplus sector. How do you reconcile this statement with the fact that many households take out loans to finance purchases that exceed their incomes?
4. Explain carefully and clearly why financial institutions such as commercial banks and savings and loan associations are called *financial intermediaries.*
5. Consider your own saving/borrowing habits over the past three years. How have you participated in the financial markets?

28

Banks and Money Creation

As you read and study this chapter, you will learn:

- the role of bank reserves in the process of check clearing
- why the size of these reserves is a crucial determinant of the size of the money supply
- how the banking system creates deposit money on the basis of its reserves
- how the Federal Reserve Bank controls the size of the money supply

Since ancient Greece, philosophers have told us that we can't step into the same stream twice. In the 20th century, physicists have pointed out that we can't do it even once. The most cautious step disturbs the stream, leaving it different from what it was before we entered the water.

During the two centuries of industrial capitalism, economic life has changed far more rapidly than in any other historical era. When most people envision this change, they think of innovations in the *techniques* of production—mechanization, assembly lines, and robotics. But *institutional* change is just as striking. We structure economic life very differently from the way we did a century or two ago, and we express this new structure in new forms of social organization.

During the late 1960s and the 1970s, rapid rates of inflation and high interest rates produced far-reaching changes in the banking industry. Old practices became outmoded, and new ones arose to replace them. The climax of this transformation

occurred in 1980, when Congress rewrote the fundamental laws governing the regulation of the banking system and redefined the powers of the ***Federal Reserve Bank.*** Since 1913, the "Fed" has been the principal agency for regulating banking in this country. At first, it was a loose confederation of regional banks, whose main functions were to provide a growing paper currency and to make loans to private banks in financial trouble. During the Great Depression, it failed miserably at the latter task. As a consequence, its powers were gradually enlarged and its structure became increasingly centralized. By the late 1950s, it had become a genuine central bank, holding deposits belonging to commercial banks, making loans to them, and buying and selling government securities on the open market. The main purpose of these three activities is to control the size of the money supply, although the Fed also continues to supply currency and ward off bank failures. The conduct of its day-to-day business is independent of the President and Congress, although its responsibilities are prescribed by legislation.

To understand how the banking system works, and how the money supply is controlled, you have to understand the functions of the Federal Reserve and know more about banking institutions than the previous chapter told you. The first section of this chapter describes how banks and the Fed fit together; the second explains money creation; and the third discusses how the Federal Reserve controls the money supply.

The institutions of the banking system

Commercial banks

You already know that ***commercial banks*** are financial intermediaries. They accept deposits from households, firms, governments, foreigners, and other financial institutions. They also make loans and buy negotiable debts. In these respects, they are like other financial intermediaries. What distinguishes the commercial banks from the others is the extent to which they take checking deposits. At the end of 1981, demand deposits made up about 20 percent of all deposits in commercial banks, and commercial bank demand deposits were about three fourths of all checking account money.

Commercial banks are classified into two sets of legal categories. The first divides them according to whether their charters were issued by the federal government or by one of the state governments. A *charter* is essentially a permit to be in the banking business. About two thirds of the banks currently in operation have state charters, and one third have federal charters. However, the federally chartered banks are larger on average than the state banks and have a bit more than half of all bank assets.

A more important distinction is whether a bank is *a member of the Federal Reserve System.* All federally chartered banks must belong to the system. State-chartered banks may belong if their directors decide membership is to their advantage, and some do. All told, only about 40 percent of banks belong, but they hold about 60 percent of all commercial bank deposits. Member bank demand deposits are about 40 percent of the M1 money supply, and total member bank deposits are about 40 percent of M2.

State banks are regulated by the states that grant their charters. Federally chartered banks are regulated by the office of the Comptroller of the Currency. Regulations govern what the banks must do for their depositors and what kinds of assets they can invest in. The details need not concern us. Their main purpose is to assure the financial soundness of the banking

system and the safety of deposits. This safety is further assured by the FDIC (Federal Deposit Insurance Corporation), which was established in 1933, following the bank failures of the Great Depression. It insures bank depositors against loss up to an amount that is currently $100,000 per deposit. About 98 percent of commercial banks subscribe to FDIC protection. However, because the $100,000 ceiling is considerably below the level of many business deposits, coverage is much less than 98 percent of deposits.

The distinction between member and nonmember banks was especially important before the passage of the Depository Institutions Deregulation and Control Act of 1980. Until then, member banks were required to keep funds on deposit with the Federal Reserve, but nonmember banks were not. Such funds, known as reserves or ***reserve deposits,*** are a powerful means by which the Fed can control the money supply. Since nonmembers were not subject to the Fed's reserve requirement, their share of the supply of bank money was not subject to its direct control. One of the principal provisions of the 1980 act was to extend the Fed's reserve requirement to cover nonmember commercial banks. To see what this extension means, you will need to understand what functions are performed by reserve deposits.

Check clearing, bank reserves, and reserve requirements

Suppose that you go to your bank and write a check to "cash" for $100. When you present it to the teller, she or he will give you $100 in currency. Every depositor has the right to cash a check up to the full amount of the balance in a demand deposit. Do you think that a bank has enough currency to cover all of its outstanding balances? Of course not. If it did, it could not make any loans or buy any securities. It would soon go broke because it would have no source of income to cover its costs. Instead, it keeps only a relatively small amount of currency and coin, called ***vault cash,*** to cover day-to-day differences between deposits and withdrawals of currency. This vault cash, which is only about 1 percent of all commercial banks' assets, is part of ***bank reserves.*** Vault cash holdings beyond what is needed to handle inflows and outflows are generally avoided, since they perform no useful function for the bank.

Suppose that instead of writing a check to cash, you write a check to your college bookstore. What will happen? If the bookstore keeps its checking deposit at your bank, very little. Your balance will drop, and the bookstore's balance will go up. For the bank, this is internal bookkeeping, nothing more. If the bookstore keeps its balance in a different bank in the same locality, things are a bit more complicated. Maybe your bank keeps an ***interbank deposit*** at the bookstore's bank. If it does, then when your check is deposited at the bookstore's bank, the bookstore's balance will be *credited*, and your bank's deposit will be reduced, or *debited.* Your bank loses an asset, its deposit at the bookstore's bank. But it also loses a liability, your deposit.

If neither of the two banks has a deposit in the other, the check may be handled through one or more ***correspondent banks.*** These are simply large banks that perform check-clearing and other services for smaller banks in their area. The smaller banks maintain interbank deposit accounts with their correspondents, which then transfer them from bank to bank as checks are cleared.

The sole reason for interbank deposits is to handle the clearing of checks. If a bank's credits just about balanced its debits every day, very small average interbank deposits could process a large volume of checks. Since they do not necessarily come

close to balancing, interbank deposits must cover any excess of debits relative to credits. Any bank that could not cover its checks would not remain open for long. Depositors would line up to withdraw their money, and the bank would have to close. Even though it might have large assets, if it did not have adequate ***liquidity*** to meet unexpected withdrawals—ready access to the means of payment—it would get into trouble. Clearly, a bank that wants to stay in business will keep *reserves* of vault cash and interbank deposits large enough to handle all foreseeable day-to-day swings in withdrawals and deposits. Reserves provide needed liquidity.

During the 19th century, bank failure due to insufficient reserves was a chronic problem. Much of the history of banking legislation concerns attempts by government to make banks more cautious than they would otherwise be. Some coercion is necessary because caution always wars with profitability. Funds that are tied up as reserves are not earning income for the banks' owners. A major form of bank regulation is the imposition of ***reserve requirements*** as a precondition of doing business.

Before the passage of the 1980 act, banks that were not members of the Federal Reserve System were subject only to the reserve requirements of their state regulatory agencies. State regulations usually required that banks hold between 8 and 15 percent of their demand deposits and a smaller fraction of their savings deposits as reserves. Vault cash and interbank deposits counted as reserves in all states. Most states permitted the banks they regulated to hold part of their reserves in U.S. government or state government securities. Since these securities are easy to sell, they can be quickly liquidated to cover withdrawals. Thus, they earn interest like loans, but are much more liquid. Holding reserve assets in government securities is a compromise between caution and profitability.

Federal Reserve member banks had their reserve requirements set by that agency. Recall that all banks with federal charters must belong to the system, and that some state-chartered banks do so by choice. About 75 percent of bank deposits were subject to the reserve requirements set by the Federal Reserve. The Federal Reserve Board required that its members hold their reserve deposits at the Fed itself. Table 1 illustrates the structure of requirements in effect at the end of the 1970s. As you can see, required reserves were higher on demand deposits than on time and savings deposits, and higher on banks with large deposits than on those with small deposits.

The percentage reserve requirements that the Fed imposed on member banks were not very different from those imposed by state agencies. What made the member bank reserve requirements so distinct from those of nonmember state banks was the range of assets that qualified as reserves. Federal Reserve member banks could count as legal reserves *only vault cash and deposits with the Federal Reserve.* Neither of these forms of reserve earned interest. The Federal Reserve provided a check-clearing service for its members (and for nonmembers that maintained deposits with the Fed). Deposits at the Federal Reserve were shifted from one bank's reserve to another's as checks were debited to one and credited to another. Thus, the Federal Reserve acted, and still acts, as a correspondent bank for its depositors. However, many members found it advantageous to maintain balances with private correspondents as well, to speed up check clearing. These did not count toward their required reserves. Nor did holdings of government securities. This put member banks at a competitive disadvantage relative to nonmembers, especially when government se-

Table 1 Member bank reserve requirements at the end of the 1970s

Requirements on Demand Deposits	Required Reserves as a Percent of Demand Deposits
Size of Bank's Demand Deposits	
First $ 0–$2 million	7
Next $2–$10 million	9 ½
Next $10–$100 million	11 ¾
Next $100–$400 million	12 ¾
Excess over $400 million	16 ¼
Requirements on Other Deposits	**Required Reserves as a Percent of Other Deposits**
Savings	3
Time	1–6 (depending on maturity date)

Source: *Federal Reserve Bulletin.*

curities were yielding a high return. There was thus a tendency in the 1970s for state banks to drop out of the system and for new banks to choose state over federal charters. Even though the federal banks were required to belong, the falloff in state bank participation weakened the Fed's control over the money supply. Further erosion of its control came from the development of new kinds of checking deposits in thrift institutions, which were not regulated by the Fed. At a time when the Fed was becoming increasingly aware of the role of money supply in the inflationary process, this weakening became a source of alarm. As a result, the Fed lobbied for, and Congress granted, a considerable extension of Federal Reserve control over the banking system.

The Depository Institutions Deregulation and Control Act

The 1980 banking law both simplified the structure of federal regulation and extended it to a wider range of institutions than those previously covered. Its main provisions are the following:

1. All *depository institutions* eligible for federal deposit insurance are covered by the act. In practice, this means virtually every sizable commercial bank and thrift institution in the country.
2. Depository institutions are subject to a set of uniform reserve requirements, based on the volume and nature of deposits, but not on the charter status or Federal Reserve membership of the institutions.
3. *Transactions deposits,* which include (a) all checking deposits and (b) any other deposits from which funds can be transferred to third parties by phone or written order more than three times a month, are subject to a reserve requirement. So are time deposits belonging to businesses and institutions, but not those belonging to individuals. Savings deposits (other than those that fall into the transactions category) are not subject to reserve requirements.
4. The reserve requirements range from 3 to 12 percent of the value of transactions deposits, depending on the size of

the institution, and from 0 to 3 percent of nonpersonal time deposits, depending on the maturity date of the deposit.

5. The Board of Governors of the Fed, its policymaking body, is empowered to impose a supplemental reserve requirement of up to 4 percent of transactions deposits.
6. Member banks must hold their reserve deposits in an account at the Fed. A nonmember institution may hold deposits at the Fed, at a depository institution that has an account at the Fed, or (if it is a thrift institution) at one of two federal agencies that regulate thrift institutions.
7. All depository institutions are eligible to borrow reserves from the Fed. Those that hold deposits at the Fed may use its check-clearing and currency-supply services.
8. The provisions of the act are to become effective gradually. They will be fully binding on member banks by 1984, and on other depository institutions by 1988.

For member banks, the new law meant no significant changes other than a restructuring of the schedule of reserve requirements. Its main effect was to extend to nonmember banks and thrift institutions those Federal Reserve regulations and services that were previously confined to the system's commercial bank members.

Commercial bank assets and liabilities

Commercial banks are the dominant suppliers of bank money, particularly demand deposits and large time deposits. They supply about 60 percent of the deposit portion of M2. Although they are gradually losing this dominance, they are likely to maintain it in the near future. When you think of deposit money, the best concrete example is the commercial bank deposit.

Table 2 provides a snapshot of commercial banking in the United States in late 1981. It is a balance sheet. Note that it is dated for a specific day rather than covering a span of time. This is because a balance sheet shows assets, liabilities, and net worth at a given moment, not flows of income over a period of time.

The first three items under "Assets"—vault cash, reserve deposits at the Fed, and interbank deposits—cover the *monetary assets* that banks use directly when honoring checks and other withdrawals. You

Table 2 Assets and liabilities of domestically chartered commercial banks, October 28, 1981 ($ billions)

Assets		Liabilities and Net Worth	
Vault cash	19.8	Demand deposits	323.9
Reserve deposits at the Federal Reserve	25.3	Time deposits	638.8
Interbank deposits	54.1	Savings deposits	214.9
Loans	901.0	Borrowings	173.3
U.S. Treasury securities	114.0		
Other securities	224.3		
Other assets	226.8	Other liabilities and net worth	214.4
Total	1565.3	Total	1565.3

Source: *Federal Reserve Bulletin.*

Table 3 ***Assets and liabilities of the Federal Reserve, October 28, 1981 ($ billions)***

Assets		Liabilities and Net Worth	
U.S. government securities	121.5		
Loans to depository institutions	1.9	Federal Reserve notes	125.7
International reserves of the		Reserves of depository institutions	26.1
United States	8.9	U.S. Treasury general account	2.8
Other assets	34.0	Other liabilities and net worth	11.7
Total	166.3	Total	166.3

Source: *Federal Reserve Bulletin.*

may be disappointed to see how small they are after all the fuss that has been made over them—only about 6½ percent of total assets. Don't let this deceive you. There is more fuss to come.

The next three items under "Assets" cover the *earning assets* of the commercial banks—loans and securities. These are the banks' bread and butter, the source of nearly all their gross income. "Other assets" include funds tied up in the check-clearing process, and in the buildings, computers, and potted palms that are needed to operate banks.

On the liabilities side are various kinds of *deposits:* demand, time, and savings. These make up the supply of commercial bank money. "Borrowings" are simply funds raised on the market to finance a portion of the banks' operations. "Other liabilities" are mainly outstanding checks that have not yet cleared. "Net worth" is a balancing item that measures approximately what would be left over for the banks' owners if all assets were sold and the deposits and other liabilities paid off.

This composite balance sheet reflects commercial banks' role as financial intermediaries. The great bulk of their assets are loans and securities. Most are financed through deposits, a form of liability. To use a term from financial analysis, banks are *highly levered*—only a small portion of their assets is financed with the equity capital of their owners. Although banks seem to be skating on thin ice with their depositors' money in their pockets, they are really quite secure. In part, this is because they deal in large numbers and can therefore predict inflows and outflows fairly well. In part, it reflects deposit insurance and the caution imposed on banks by reserve requirements and other regulations. An undeniable third element in the security of banks and their depositors is the chartering and regulation of banks, which restricts competition and carves up the market into pieces large and stable enough to ensure that few banks will fail.

Federal Reserve assets and liabilities

The Federal Reserve also has a balance sheet. Like that of the commercial banks, it is a snapshot of the bank's major activities. Table 3 presents such a snapshot, dated to coincide with Table 2. It mirrors the Fed's role in regulating and serving the banking system and its other responsibilities.

As you can see, the major assets of the Fed are U.S. government securities. These are a part of the national debt, which is built up as the federal government runs budget deficits. The Federal Reserve routinely buys negotiable government securities when it wants to expand the money

supply, and sells them when it wants to contract the money supply.

The second item on the asset side of the account consists of outstanding loans to depository institutions. Such loans provide temporary reserves to banks that would otherwise have difficulty meeting their reserve requirements.

The third item consists of Fed holdings of deposits in other countries. These are used to intervene in the international currency markets. One of the Fed's responsibilities is to stabilize the international value of the dollar when it is threatened by speculation. You will learn about this in a later chapter.

The Fed's "Other Assets" are its unsettled accounts with the banks whose check clearing it handles and its physical facilities. Like the commercial banks, it has its share of buildings, computers, and potted palms. Indeed, since the Fed has a large income from the government securities it owns, it boasts some of the finest potted palms in the banking business.

The main entry on the liability side is the outstanding supply of Federal Reserve notes, the country's paper currency. The second most important item is the reserve deposits of depository institutions. Most of this is the $25.3 billion of commercial bank reserves that appear in Table 2. The rest belongs to other depository institutions. The $2.8 billion deposit belonging to the U.S. Treasury is the Treasury's bank account, in which it deposits tax receipts and from which it pays its bills. The "Other Liabilities" are dominated by bookkeeping items connected with check clearings. There is also a small "Net Worth" entry. **Although the Federal Reserve was created by Congress and is largely governed by officials appointed by the President and confirmed by the Senate, it is an independent institution *owned by the member banks.* The net worth represents the accounting value of their ownership.**

The creation of bank money

Now that you know the names and numbers of the major players, you are ready to follow the process of money creation. Bank money—the supply of checking, time, and savings deposits—is created by the lending of depository institutions. Any such institution can create money, although the details of the process differ somewhat between commercial banks and thrift institutions.

Bank lending and deposit creation

Suppose you borrow from a commercial bank. The bank will write you a check drawn on itself. If you deposit the check in your account at the same bank, the bank credits the amount of the loan to your account. If you cash it, the lender bank loses vault cash, part of its reserves. If you deposit the check in another bank, the lender loses reserve deposits to the bank where you deposit the check. But even if you deposit the check in the bank that makes the loan, that bank can count on losing reserves when you start writing checks for the things that made you borrow the money in the first place. Odds are that few if any of these checks will be deposited in the lender bank. **Therefore, *banks do not lend money unless they can afford an outflow of reserves.***

Remember that *reserve requirements* compel banks to hold vault cash and reserve deposits equal to a percentage of their deposits. But any bank that holds more than the required reserves can make loans because it can afford to lose those reserves when the borrower spends the money. A bank with such a reserve position is said to have ***excess reserves.*** The amount of excess reserves is given by the following relationship:

$$\text{Excess reserves} = \text{Actual reserves} - \text{Required reserves.}$$

Excess reserves form the basis for monetary expansion through bank lending.

To see how this works, suppose that the owners of the New York Knickerbockers basketball club apply to the Manufacturers Hanover Trust Company, a commercial bank in New York City, for a $1 million loan to pay a bonus to a free-agent player. Maybe he is a "strong forward." (They all look strong to the rest of us, but apparently some of them are stronger than others.) If the Manufacturers Hanover is ***loaned up***—without excess reserves—it cannot make the loan unless it can somehow get more reserves. If it has excess reserves, it can accommodate the Knicks' management because it can afford to lose reserves. Suppose that the loan is made, and that after the bonus is paid, the player sends the $1 million check to his mother in Michigan, who deposits it in the Bank of the Commonwealth in Detroit. After all the checks have cleared, what has happened to the various accounts involved?

The answer may be seen in Figure 1, which is made up of a series of related "T-accounts." These handy little devices (which look like the letter T) are very useful in tracing asset and liability *changes* for a single institution on a single transaction. Items that do not change do not appear. Because an asset transaction is an exchange of equivalents, each T-account must balance: The sum of changes on the left, taking account of sign, must equal the sum of changes on the right.

The top account shows that Manufacturers Hanover lost reserves, as it expected when it made the loan, which balances the lost reserves in the bank's assets. Both Manufacturers Hanover and the Bank of the Commonwealth keep their reserves at the Fed. The Fed account shows that the reserves lost by Manufacturers Hanover were transferred to the Bank of the Commonwealth in the check-clearing process. The Commonwealth's account registered

Manufacturers Hanover Trust Company

Assets	Liabilities
Reserve deposits at the Fed −1 mil.	No change
Loans +1 mil.	

Federal Reserve Bank

Assets	Liabilities
No change	Reserves of Manufacturers Hanover −1 mil.
	Reserves of Bank of the Commonwealth +1 mil.

Bank of the Commonwealth

Assets	Liabilities
Reserve deposits at the Fed +1 mil.	Mother's demand deposit +1 mil.

Figure 1 The effects of a loan

When Manufacturers Hanover makes a loan, and the proceeds are deposited elsewhere, its reserve assets go down, and its loan assets go up. The Fed transfers reserves from one bank to another. Since bank reserves are liabilities of the Fed, this transfer appears on the right-hand side of the Fed's balance sheet. If the loan proceeds are deposited in the Bank of the Commonwealth, its reserve assets go up, but so do its deposit liabilities.

an increase in both reserves and deposits. There is nothing to show that the Knicks got a new strong forward with a rich mother, but her $1 million in newly created money is there among the liabilities of the Bank of the Commonwealth.

Multiple deposit creation

This increase in the money supply was achieved by mobilizing unneeded reserves. All money creation by the banks themselves, without any help from the Fed,

takes place in this way. If a bank has excess reserves, it may increase its loans or buy securities. In either case, it creates new deposits. Or it may lend reserves to other banks through the market for ***federal funds,*** which is simply a loan market for reserves. The borrower on the federal funds market can then lend to customers or buy securities.

It may strike you as mildly interesting that a commercial bank loan creates money, but the process is far more interesting than we have let on so far. If you look back at Figure 1, you will see that when the Manufacturers Hanover Bank lost reserves, as it expected to when it made the loan, the reserves did not disappear. There they sit in the reserve account of the Bank of the Commonwealth, providing the potential for further loans and money creation. If the Commonwealth could make a $1 million loan, it, too, would lose $1 million in reserves, but they would show up somewhere else, giving some other bank the potential for lending and creating still more deposits. It almost looks as though the $1 million of excess reserves could form the basis of an infinite chain of money expansion, limited only by the public's willingness to borrow and the rapidity with which the excess reserves circulate.

In fact, this is not how things work. If you look back at Figure 1 again, you will see that besides gaining $1 million in reserves, the Bank of the Commonwealth has also gained $1 million in deposits. Since its required reserves are calculated as a fraction of deposits, they must, therefore, have risen. Thus, the Commonwealth cannot afford to lose the whole $1 million in reserves. Its lending capacity has only increased by some fraction of $1 million. Since required reserves increase as deposits are created, a limit is placed on the banking system's capacity to create new deposit money.

To understand how banks create money, it will help you to follow the process of deposit creation step by step. Like any complicated social process, though, deposit creation has both essential properties and incidental details. To focus on the essential properties, we will make several simplifying assumptions, then remove some of them later to fill in the incidental details. Suppose that:

1. All money consists of bank deposits, with no circulating currency or coin.
2. Banks do not accept time or savings deposits, only demand deposits.
3. All banks keep their reserves at the Federal Reserve Bank.
4. All deposits are subject to a flat 20 percent reserve requirement.
5. Lending is so profitable that no bank voluntarily holds excess reserves.
6. Loan demand is so strong that no bank has trouble finding borrowers.
7. All banks are initially loaned up, without excess reserves, except for Manufacturers Hanover, which has $1 million in reserves beyond the required 20 percent of its deposits.

Begin the process of money expansion with the Manufacturers Hanover loan to the New York Knicks. Figure 1's T-accounts have shown you the mechanics of the check clearing, reserve transfer, and deposit creation involved. Think of it as the first of a series of "rounds" of ***multiple deposit creation,*** the top row in the body of Table 4. The initial $1 million of excess reserves gives rise to an equal increase in bank loans and deposit money. Because the new money takes the form of bank deposits, required reserves of the banking system go up, in this case by $200,000, or 20 percent of the deposit increase.

The second round begins with the Commonwealth in possession of the $1

Table 4 Reserves and deposit creation

Round	Initial Excess Reserves	Increase in Loans	Deposit Money Created	Rise in Required Reserves
1	$1,000,000	$1,000,000	$1,000,000	$ 200,000
2	800,000	800,000	800,000	160,000
3	640,000	640,000	640,000	128,000
4	512,000	512,000	512,000	102,400
5	409,600	409,600	409,600	81,920
•	•	•	•	•
•	•	•	•	•
•	•	•	•	•
•	•	•	•	•
•	•	•	•	•
	0	0	0	0
Sum	Not relevant	5,000,000	5,000,000	1,000,000

In each round of expansion, the banks make loans equal to their excess reserves, creating equal amounts of new money. If required reserves rise by 20 percent of the increase in deposit money, in each round, the reserve excess is only 80 percent as large as it was in the previous round.

The rounds continue until required reserves increase by the amount of the initial reserve excess. Since the reserve ratio is 20 percent, it takes a $5 million rise in bank deposits to raise required reserves by $1 million. Thus $1 million in excess reserves permits a $5 million increase in the money supply.

million of additional reserves, $200,000 of which it must keep because of its larger deposits, and $800,000 of which it can afford to lose without violating the reserve requirement. If it writes an $800,000 check to a borrower, it gets a profitable earning asset of equal value, but it loses reserves when the borrower's check is deposited elsewhere. Other banks get the reserves, and equivalent new deposits. Their required reserves go up by 20 percent of the deposit increase, leaving them with excess reserves equal to 80 percent of the reserves they gain. This increases their lending capacity, and the process carries on to another round—but the rounds get successively smaller.

There are two ways of figuring the total money supply increase that comes from an initial $1 million in excess reserves. One way is to add up the money created on the various rounds. You will notice in Table 4 that (1) in each round, the money created equals the excess reserves at the beginning of the round; and (2) that the excess reserves drop by 20 percent in each round, since required reserves rise by 20 percent of the money created. This means that the sequence of amounts created can be summed according to the formula for a geometric series in which every term is 80 percent of the preceding term. Thus, according to the rounds of Table 4, the money creation is:

$$\text{Total money creation} = \$1{,}000{,}000 + 800{,}000 + 640{,}000 + 512{,}000 + \ldots$$

But the money creation is also given by:

$$\begin{aligned}\text{Total money creation} &= \$1{,}000{,}000\,(1 + .8 + .8^2 + .8^3 + \ldots) \\ &= \$1{,}000{,}000\,\frac{(1)}{(1-.8)} \\ &= \$1{,}000{,}000\,\frac{(1)}{(.2)} \\ &= \$1{,}000{,}000 \times 5 \\ &= \$5 \text{ million.}\end{aligned}$$

The money created is 5 times the initial amount of excess reserves, where 5 equals *the reciprocal of the reserve requirement.* In form, this resembles the calculation that gives the GNP multiplier. The ratio of money created to excess reserves is sometimes called the ***bank deposit multiplier.*** If you want to use this term, go right ahead. Just be sure to remember how it differs from the GNP multiplier.

The second way of figuring the money supply increase is to think carefully about what happens to the initial excess reserves. These reserves never leave the banking system. They are not loaned to individuals; only banks can own reserve deposits. Yet, eventually, the *excess* reserves disappear because *required* reserves rise to absorb them. The expansion process ends when the additional reserves required by the new deposits just balance the initial excess. Then the banking system is loaned up, and there are no excess reserves that a bank can afford to lose. So:

$$\text{New money created} \times \text{Required reserve ratio} = \text{Initial excess reserves}$$

or

$$\frac{\text{New money created}}{\text{Initial excess reserves}} = \frac{1}{\text{Required reserve ratio.}}$$

In terms of our numbers, this would be:

$$\frac{\$5 \text{ million}}{\$1 \text{ million}} = \frac{1}{.2} = 5$$

where the .2 is the 20 percent reserve requirement.

Deposit contraction

All of this also works in reverse, of course. If Manufacturers Hanover is short of reserves instead of having an excess, it must build them up to the legal minimum. To do this, it could borrow reserves from the Fed, but this would just delay the day of reckoning. It could also borrow reserves on the federal funds market, but if other banks are loaned up, there are no reserves to borrow. The only thing it can then do to get reserves is to reduce its income-earning assets—to sell some of its securities or refuse to renew a loan that is coming due. Of course, it loses income-earning assets when it does this, but it gains reserve deposits when the check that buys the securities or pays off the loan clears through the Fed.

To be concrete, suppose that because Manufacturers Hanover is short $1 million in reserves, it refuses to renew its loan to the Knicks when it comes due. To pay off the loan, the Knicks have to write a check against their gate receipts, which they keep in the Chemical Corn Exchange Bank. (A real bank: These long bizarre-sounding names result from mergers of banks with shorter names. This name has, in fact, been shortened to the Chemical Bank, which is still odd enough.) When the check clears, the Fed transfers reserves to Manufacturers, solving its ***reserve deficiency.*** Unfortunately, the Chemical Bank has inadvertently acquired a similar problem, since it has lost reserves. The results of these transactions are shown in Figure 2. However, it has also lost deposits from the Knickerbockers' account, so that its required reserves have dropped. Therefore, its pickle is not quite so bitter as the one that Manufacturers had to bite.

Table 5 shows the successive rounds of deposit contraction that are triggered by Manufacturers' initial reserve deficiency. Round 1 presents the actions of Manufacturers and their direct implications for the Chemical Bank. Manufacturers' $1 million reserve deficiency causes it to contract its loans by $1 million. The Chemical Bank loses $1 million in deposits and reserves as the Knicks pay off their loan. Its required reserves drop by $200,000, but it still has a reserve deficiency of $800,000 at the beginning of Round 2. It must contract its loans or security holdings by the same

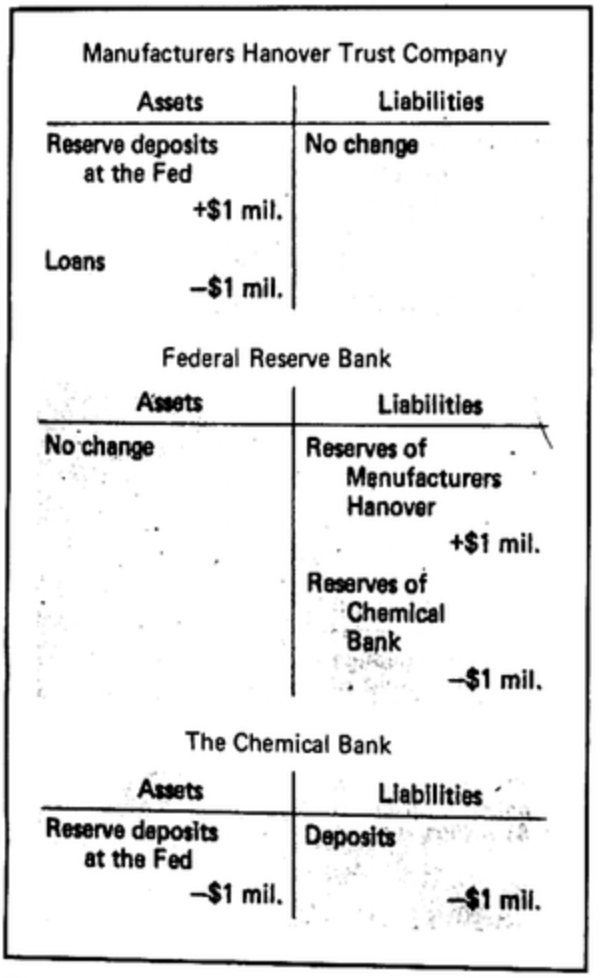


Figure 2 The effects of retiring a loan

When Manufacturers Hanover calls in a loan, and it is paid off with a check from another bank, its reserve assets go up, and its loan assets go down. The Fed transfers reserves from one bank to another. When the check that pays off the loan is debited against the Chemical Bank's reserves and against the Knicks' account at that bank, its reserve assets go down, but so do its deposit liabilities.

amount. This wipes out $800,000 of deposits and reserves somewhere else in the banking system and triggers another round. The rounds continue until $5 million of deposit money has been destroyed. After this has happened, required reserves have shrunk by $1 million, and the system as a whole no longer has a reserve deficiency.

This example of deposit contraction is perfectly symmetrical with the example of deposit creation. The numbers in Tables 4 and 5 are the same; only the labels are different. And both processes embody an important but subtle truth: **If banks try to get reserves by contracting loans, or to get rid of unwanted reserves by expanding loans, they cannot collectively succeed because *banks do not control the quantity of reserves. Their collective attempts to gain or lose reserves only change the supply of deposit money until required reserves equal the actual reserves that are available.***

Who, then, controls the quantity of reserves? As the next major section shows, it is none other than the Fed. But before turning to the Fed's control over bank reserves, we will replace some of the complications of deposit creation that have been ignored so far in the interest of simplicity and clarity.

Some complications

The first complication is that deposit expansion is not mandatory. Banks with excess reserves don't have to increase their earning assets. And when they do try to expand their loans or security holdings, they may not find borrowers. The whole expansion process depends on both the banks' willingness to lend and the public's willingness to borrow. During the Great Depression, the banking system had large excess reserves for an extended period. This probably resulted both from caution by banks and from pessimism by people who in more prosperous times would have financed more loans.

The contraction that follows from a reserve deficiency *is* mandatory, however. Banks with a reserve shortfall pay a penalty to the Fed unless they make it up quickly. This means that the quantity of reserves sets an upper limit to the supply of bank money. Banks may not go beyond their loan capacity, though they may stay within it. Thus, expansion and contraction are only symmetrical under conditions that make it possible and profitable for banks always to stay loaned up.

Table 5 ***Reserves and deposit contraction***

Round	Initial Reserve Deficiency	Contraction in Loans or Security Holdings	Deposit Money Destroyed	Drop in Required Reserves
1	$1,000,000	$1,000,000	$1,000,000	$ 200,000
2	800,000	800,000	800,000	160,000
3	640,000	640,000	640,000	128,000
4	512,000	512,000	512,000	102,400
5	409,600	409,600	409,600	81,920
•	•	•	•	•
•	•	•	•	•
•	•	•	•	•
•	•	•	•	•
•	•	•	•	•
	0	0	0	0
Sum	Not relevant	5,000,000	5,000,000	1,000,000

In each round of contraction, the banks contract their loans or security holdings by an amount equal to their reserve deficiency, destroying an equal amount of deposit money. Required reserves drop by 20 percent of the drop in deposits, so that in each round, the reserve deficiency is only 80 percent as large as it was in the previous round.
The rounds continue until the drop in required reserves equals the initial reserve deficiency. Since the reserve ratio is 20 percent, it takes a $5 million drop in the money supply to lower required reserves by $1 million. Thus a $1 million reserve deficiency leads to a $5 million decrease in the money supply.

The presence of excess reserves modifies the contraction of deposits in the same way that the possibility of keeping excess reserves modifies expansion. If the banking system as a whole has excess reserves, then a bank may borrow reserves on the federal funds market. Therefore, any bank with a deficiency can remedy it with federal funds, without having to sell income-earning assets or to get temporary help by borrowing from the Fed. Both multiple expansion and contraction are therefore smaller when banks do not keep loaned up.

A second complication comes from the other component of the M1 money supply, currency. When loan proceeds are spent, some of the money created circulates as increased currency. The reason for this *cash drain* is simply that people want to hold currency; they tend to keep some rough balance between their deposit and currency holdings. Banks are obliged to accommodate this behavior. With an expanding demand for currency, banks find themselves losing vault cash, which they ordinarily replenish with new currency from the Fed. To get this new currency, they must give up reserve deposits. The banks' action imitates that of their depositors. When you cash a check, you get currency and give up demand deposits. When the bank replaces the lost vault cash, it gets currency and gives up reserve deposits. This is how currency gets into circulation.

Because currency holdings increase along with the supply of deposit money, excess reserves have smaller expansionary potential than they otherwise would. Suppose, for example, that the Knicks' new forward takes half his $1 million bonus in cash and puts it under his king-sized mattress, sending a check for the other half to his mother. When she deposits it in a bank, it provides the basis for a $2.5 million expansion in bank deposits, but the total money created is only $3 million—the $500,000 in currency under the mattress

and $2.5 million in deposits—rather than the $5 million that was created when the entire reserve excess stayed in the banking system. The reason for the smaller expansion in the money supply is that withdrawal of currency deprives the banking system of reserves that otherwise would lead to multiple deposit creation. Indeed, if the forward takes the entire loan in cash, the banking system's $1 million of excess reserves goes under his mattress, and there is no deposit creation at all.

Remember that banks may count both vault cash and reserve deposits to meet the requirements imposed by the Fed. When they lose vault cash, they lose reserves, even if they don't have to get new currency from the Fed. This loss of actual reserves has the same effect on bank credit expansion as does a rise in required reserves. Successive rounds of money creation get smaller because some of the lending capacity is drained away into circulation.

Finally, there is a host of minor complications stemming from the existence of time and savings deposits and varying reserve ratios on banks of differing sizes. These mess up the arithmetic of the deposit multiplier considerably, but do not affect the principle. As long as the fractional reserve requirements on all deposits are less than 1, any new deposit creates some potential for a loan. As long as the average reserve requirement is greater than zero, the successive rounds of expansion must get smaller and smaller. Eventually they die out.

The Federal Reserve and money creation

As you already know, the Federal Reserve sets reserve requirements for the banking system, within limits imposed by congressional legislation. It also lends reserves to banks that have a shortfall, and shifts reserves from one bank to another as checks are cleared. The other major function of the Fed that has only been touched upon so far is its creation and destruction of bank reserves themselves.

The activities it carries out to create and destroy reserves are called ***open market operations.*** These are Federal Reserve purchases and sales of U.S. government securities on the financial markets. When it buys securities, for example, the Fed creates reserves for the banking system equal in value to its purchases. When it sells securities, it reduces bank reserves by the same amount. Having just gone through the implications of reserve excesses and deficiencies, you can appreciate that these changes in reserves have a multiplied impact on the money supply. This is exactly what they are designed to do. Open market operations are deliberate attempts to control the money supply.

Before going into the mechanics of open market operations, though, you might consider why the Fed or any other government agency would want to control the money supply. The reason is simple: Bankers, economists, utopians, and outright cranks all agree that a country with an intelligently managed money supply is more stable than one with a chaotic money supply. In managing the money supply, the Fed is just doing what everyone agrees upon in principle. In practice, of course, there is little agreement on specifics. Programs for money management range from rigid rules on growth to leaving money creation up to the caprices of the world gold supply. The Fed thinks it can do better than either of these alternatives. When you study stabilization policy in the next few chapters, you can decide whether the Fed is right.

The monetary base

Look back at Table 3, which presents a Federal Reserve balance sheet for October

28, 1981. On the liability side of the ledger, you can see $125.7 billion in Federal Reserve notes (i.e., currency) outstanding and $26.1 billion in bank reserves. Of these reserve deposits, $1.9 billion consisted of bank borrowings from the Fed (see the asset side), so that its net reserve debt to the banking system was $24.2 billion. The sum of the outstanding currency and net reserve deposits was the ***monetary base*** or stock of ***high-powered money*** on that particular date—$149.9 billion. Part of the monetary base belonged to the banking system—about $45 billion in reserve deposits and vault cash. The remaining $105 billion—consisting of all the currency not held in the vaults of the banking system—was in the hands of the general public.

Why is it called the "base" of the money supply, and in what sense is it "high powered?" Obviously, the quantity of reserves owned by the banking system is the base of an inverted pyramid of deposit money that the banking system can supply even if the Fed does not lend it any further reserves. These net reserves are the banks' unborrowed reserve deposits and vault cash. For every dollar of such reserves, the banking system can create several dollars of bank money through the process we have just explained. This makes these reserves high powered. The currency outside banks in the hands of the public also has the *potential* for being high powered. If the public deposits it in banks, it, too, joins the stock of bank reserves and forms a basis for multiple deposit creation.

Open market operations

The national debt is more than $1 trillion. In tangible form, this debt consists of securities of varying maturities. Whenever the U.S. government runs a budget deficit, it finances this deficit by selling more securities. Some of the debt is owned by government agencies, including Social Security and the Federal Reserve. But most of the securities are owned by private investors.

The Fed holds about 12 to 15 percent of the federal debt. Unlike private investors, it buys, holds, and sells its securities to control the money supply, not to make an income. To see how this works, we will discuss the effects of a hypothetical Federal Reserve purchase.

Suppose that the Fed decides to buy $10 million worth of one-year Treasury notes at the same time that the Exxon Corporation wants to sell an equal volume of such notes, perhaps to finance capital investment. The transaction is made through a broker that specializes in Treasury securities. The Fed gets the notes, and Exxon gets a check drawn on the Fed. It deposits this check in an account at Citibank in New York, where it keeps an account. When the check clears, Citibank's reserve account at the Fed is credited with $10 million, and Exxon's account at Citibank is credited with the same amount. The balance sheet effects of this transaction are listed in Figure 3 and diagrammed in Figure 4. The most important thing about this transaction is that it raises the quantity of bank reserves—part of the monetary base—by the amount of the Federal Reserve's purchase. It also directly increases Exxon's holdings of deposit money by $10 million. When Exxon spends this money on capital goods, both deposits and reserves will be dispersed throughout the banking system. Every bank that ends up with a part of the $10 million will have excess reserves, since its actual reserves will have grown by the same amount as its deposits, but its required reserves will have grown only by some fraction of the deposit increase. Thus, the banking system as a whole will be able to create more money.

This, of course, is what the Fed has in mind when it buys government securities in the first place. It wants to increase the

Federal Reserve

Assets	Liabilities
Government securities +$10 mil.	Reserves of Citibank +$10 mil.

Exxon

Assets	Liabilities
Government securities −$10 mil.	No change
Deposits at Citibank +$10 mil.	

Citibank

Assets	Liabilities
Reserves at the Fed +$10 mil.	Deposits of Exxon +$10 mil.

Figure 3 The effects of a Federal Reserve open market purchase

When the Federal Reserve buys $10 million worth of government securities from Exxon, it directly creates $10 million of bank money. It also furnishes $10 million of reserves to the banking system, which provides the basis for further deposit creation.

monetary base to increase the money supply indirectly because it thinks that the best interests of the economy will be served by a larger money supply.

When it wants to decrease the money supply, it sells securities on the open market. When the purchasers' checks clear, the banking system loses reserves, again part of the monetary base. The reduction in the monetary base sets off a multiple contraction in the money supply.

Changes in reserve requirements and the Fed's lending rate

Open market purchases and sales of government securities are the most frequently used means that the Fed has to influence the money supply. Sometimes, however, it wants to make a big change in a hurry. If it wants to increase the money supply quickly, it may decide to lower required reserve ratios. This does not create additional reserves, but it immediately creates

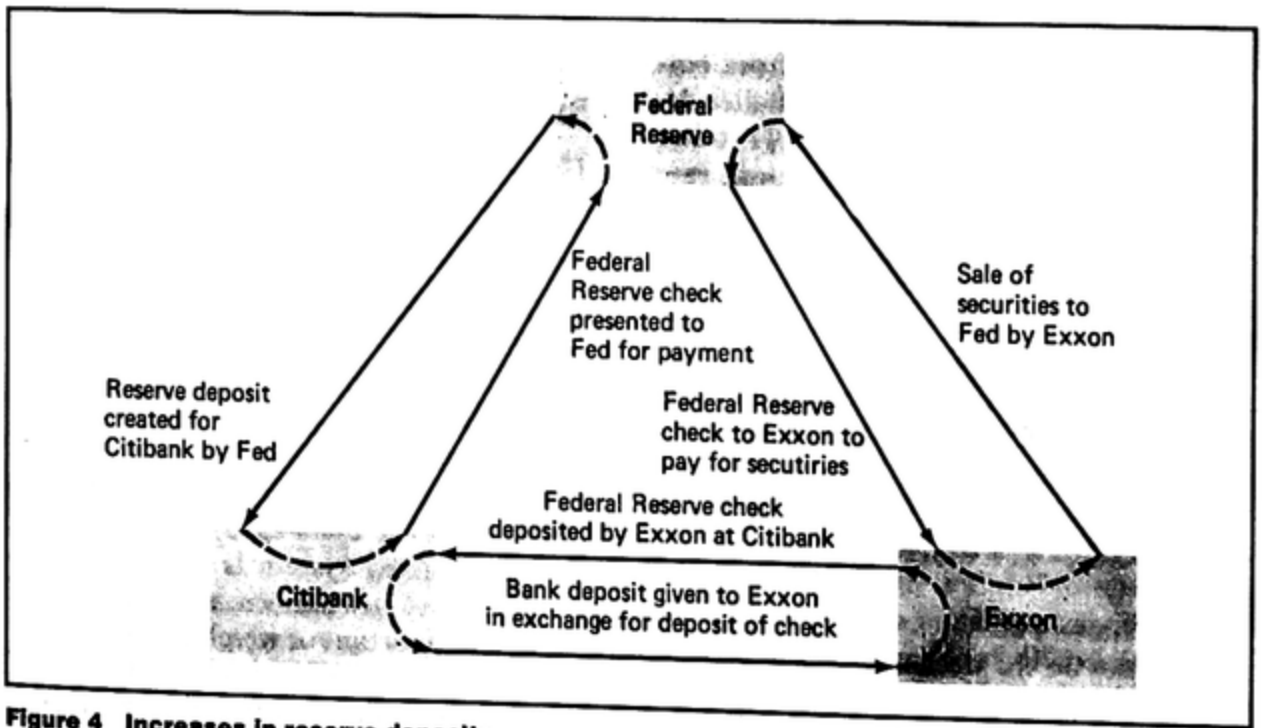


Figure 4 Increases in reserve deposits

The transactions shown in Figure 3's T-accounts involve a three-way swap of assets and liabilities among the Fed, Exxon, and Citibank.

excess reserves for *all* banks by reducing required reserves. This permits additional ***deposit creation.*** Such a reduction in required reserves raises the ratio of the potential money supply to the monetary base. In effect, it raises the deposit multiplier. An increase in the reserve requirement similarly reduces both excess reserves and the deposit multiplier, forcing the money supply to contract.

Changes in the interest rate that the Fed charges when it lends reserves to its members have more subtle effects. Suppose that a member bank has a reserve shortfall and the rest of the banking system is loaned up. It may choose to reduce its outstanding loans to get additional reserves, or it may borrow from the Fed. If it does the first, there will be a multiple contraction in the money supply. If it borrows from the Fed, there will not. When the ***Federal Reserve's lending rate*** (also called the ***discount rate)*** is relatively high, the bank will choose to contract its loans. When the rate is relatively low, the bank will borrow from the Fed. Thus, a higher discount rate discourages banks from borrowing reserves, and leads to a smaller deposit multiplier and money supply, other things being equal. A lower discount rate encourages banks to borrow reserves and leads to a larger money supply.

The Fed's operations as a whole

The Federal Reserve, then, has three main levers of control over the money supply: its open market operations, its lending rate, and its required reserve ratio. Each of these operates in a different way:

1. Purchases and sales of government securities on the open market directly affect the monetary base by increasing or decreasing the net reserve position of the banking system.
2. Increases or decreases in the Fed's lending rate affect total bank reserves by discouraging or encouraging banks to borrow reserves.
3. Decreases or increases in required reserve ratios affect excess reserves by lowering or raising the amount of reserves required per dollar of deposits.

All of these methods operate indirectly on the money supply. Indeed, the Fed has *no* direct control over the money supply itself. Through its open market operations, it controls the monetary base. Through its reserve lending and its power to control reserve requirements, the Fed sets some of the terms on which banks and the public determine the volume of bank lending, and therefore how large the money supply will be relative to the monetary base. But overall, the money supply is determined not by the Fed alone, but by a three-way interaction among the Fed, the banking system, and the general public.

Summary

This chapter contains a mixture of theory and institutional background material. You should remember the following:

1. The American money supply is largely created by banks, particularly commercial banks. This is true not only of deposit money, which is the liability of the banking system, but also of currency, which gets into circulation through the banking system.
2. The banking system is regulated by a variety of governmental agencies, the most important of which is the Federal Reserve Bank, or Fed. The Fed requires banks to hold a portion of their assets in specified forms known as reserves.

This portion equals some fraction of the banks' deposits. The remainder of the banks' assets are mainly loans and securities, which are the sources of their earnings. Bank liabilities are mainly deposits.

3. Any bank that has reserves in excess of requirements may increase its loans and security holdings by writing checks on itself. When a borrower uses the loan proceeds to buy things, the lending bank loses both deposits and reserves. However, these deposits and reserves are transferred through the check-clearing process to other banks, which can use a portion of the additional reserves for further lending and deposit creation. The total bank money created is some multiple of the initial reserve excess. The public may, of course, take some of the increased money in currency.
4. This process also works in reverse. If a bank has a reserve deficiency, it must contract its loans, and in the process, the supply of deposit money is reduced.
5. The banking system as a whole gets additional reserves through the open market operations of the Federal Reserve. When the Fed buys U.S. government securities on the financial markets, it pays for them by writing a check on itself. When this is deposited, it increases the reserve deposits that banks hold at the Fed and helps expand the money supply.
6. This, too, works in reverse. When the Fed sells government securities, it reduces the reserves of the banking system and forces the money supply to contract.
7. Open market operations are part of the Federal Reserve's program of money control in the interest of stability, prosperity, and growth. The Fed's other principal levers of control are the power to set reserve requirements and to set the terms on which it will lend reserves to the banking system.

Key concepts

The Federal Reserve Bank (the "Fed")
Commercial banks
Reserve deposits
Vault cash
Bank reserves
Liquidity
Reserve requirements
Excess reserves
Loaned up
Federal funds
Multiple deposit creation
Bank deposit multiplier
Reserve deficiency
Open market operations
Monetary base or high-powered money
Deposit creation
Federal Reserve's lending rate or discount rate

Questions for review

1. A student in your economics class is worried. She has just read that banks do not have enough currency to cover all of their outstanding deposits. She's afraid that she'll never again see the money in her checking account. Reassure her, using the material covered in this chapter.
2. How does the Depository Institutions Deregulation and Control Act of 1980

tighten the control of the Fed over the money supply?

3. Are the following statements true or false? Explain.
   a. Since owners' equity is only a small part of a bank's assets, banks are fairly risky enterprises.
   b. Even if there is no cash drain, expansion in the money supply that results from an increase in reserves must end eventually.
   c. If banks find themselves short of required reserves, they can contract loans, and thereby increase the quantity of reserves in the banking system.
4. Define the *monetary base* (or high-powered money). How is it related to the money supply?
5. a. What are the three major methods through which the Fed can influence the money supply?
   b. Which of these methods works directly on the monetary base? Which of these methods works on the relationship between the monetary base and the money supply?
   c. How would the Fed manipulate each of these tools to decrease the money supply? Explain.

29

Money, Interest, and GNP

As you read and study this chapter, you will learn:

- what determines the level and structure of interest rates
- why the interest rate influences the demand for goods and services
- how the interest rate and GNP mutually affect each other
- how Federal Reserve policy influences GNP
- what the main differences are between monetarism and Keynesianism
- what role money plays in the inflationary process

Money is important. Nearly everyone thinks so. In this case, they happen to be right. Yet, even among economists, there is a wide range of opinion on just how much it matters. Some claim that changes in the money supply only moderately influence GNP. Others insist that monetary changes dominate the business cycle in the short run and the price level in the long run.

There is a lot of joking, most of it good natured, about the inability of economists to agree. Two old favorites are "I asked four economists and got five different answers," and "If all the economists in the country were laid end to end, they wouldn't reach a conclusion." Some of this is unfair. In a field in which experimentation is almost unheard of, it is very hard to come to conclusions that everyone supports. Some of the criticism is richly deserved, however. Economists study economics precisely because they care deeply about society. They become entangled in the major social controversies of their times in ways that can't help influencing their perceptions and therefore their conclu-

sions. Their analysis becomes a mixture of science, ideology, and politics.

To appreciate the range of disagreement possible in economics, it is instructive to compare the writings of Professors Milton Friedman of the University of Chicago and Paul Samuelson of the Massachusetts Institute of Technology. Both have authored many important papers and books. Both have written weekly economic commentaries for *Newsweek* magazine. Friedman has been for many years the principal champion of the *monetarist* position, which holds that changes in the supply of money dominate the fluctuations of GNP. Samuelson has been a major advocate and popularizer of the *Keynesian* position, which places much less emphasis on money, and much more emphasis on the impact of the federal budget. The Nobel Prize Committee has seen fit to award both of them the prize in economics. Does this signify that somehow both are right? Hardly. In a sense, it is an open recognition that controversy in economics is chronic and is likely to remain so.

You might think that differences on so obviously quantitative a matter as the importance of changes in the money supply could be resolved by looking at the facts. But facts hardly ever speak for themselves. Intricate relations cannot be measured by plotting a couple of variables on a graph. The type of research that works fairly well for the consumption function, for example, does not work at all for studying changes in the cost of credit. Complicated statistical procedures are called for. When issues are complicated, there are usually many ways of approaching them, and not much agreement on the best way to proceed. The problem of reaching agreement on the facts is all the more difficult because the different ways of organizing data often give strikingly different results. A careful person should be skeptical of all positions on the role of money that are stated with absolute certainty. Yet, the issue of the importance of money cannot be avoided, since it is obviously crucial for understanding much of the rest of macroeconomics.

The main link between the financial markets and the goods markets is the rate of interest. Causation runs both ways. Fluctuations in GNP relative to the supply of money cause the interest rate to change. Fluctuations in the interest rate cause GNP to change. This means that understanding macroeconomic events requires a simultaneous look at the financial markets and the markets for goods and services.

This chapter is divided into three main parts. The first explains how the interest rate is determined. The second tells why some of the components of planned demand respond to changes in the interest rate. The third ties the first two together.

The determination of interest rates

If you are a typical college student, your direct contact with credit and security markets is probably limited to a savings account, some charge accounts or credit cards, and maybe a student loan. If you are older, you may own your own home and have a mortgage. Maybe you are making payments on a car. But in all likelihood, you have never seen anything so exotic as a Treasury bill, a corporate debenture, or any commercial paper. Chances are, you never will. The variety of kinds of indebtedness that are routinely marketed in our economy is staggering. Unless you enter a profession that brings you into contact with the financial markets, you are not likely to encounter very many of them.

The rate of interest

All financial instruments, as loans and securities are sometimes called, yield some sort of ***rate of return.*** A rate of return is a

ratio, equal to the annual *income* returned to the owner divided by the amount of money, or *principal*, invested in it. Lenders who have funds to invest look closely at the rate of return, along with safety and liquidity, or ease of resale, in trying to decide where to put their money. When viewed by a would-be borrower, the ratio of income to invested principal is called the *cost of credit* or ***cost of funds***—the rate that must be paid to get funds. The rate of return and the cost of credit are thus the same thing viewed from the different viewpoints of the lender and the borrower. A third name for the same thing is the ***rate of interest.***

You might suppose that you would have to study many interest rates to understand how they interact with other economic variables. Fortunately, this is not true. The rates of return on nearly all financial instruments go up and down together. Competition among borrowers and lenders, particularly the large financial institutions, makes this happen. (Some concrete illustrations are given in the box accompanying this text.) This is an enormous help in figuring out how financial markets interact with the rest of the economy. You can think about changes in the whole constellation of rates of return by concentrating on a single rate of return, representative of them all. For this purpose, we will generally use the yield on three-month U.S. Treasury bills as *the rate of interest.*

Money demand, interest, and velocity

The equilibrium rate of interest is determined by the demand for and supply of money. The *supply* of money is largely controlled by the Federal Reserve, but as you know from the previous chapter, it is also influenced by the public and by financial institutions.

The *demand* for money comes from the households, firms, and governments whose transactions make up the circular flow of goods and services. It is easy to see that the demand for money is linked to the level of GNP. The higher the money value of GNP is, the higher are the money transactions that make up the circular flow, and the greater is the amount of money needed in circulation. The lower the money value of GNP, the lower are the money transactions, and the lower is the need for money. Thus, the level of income and the demand for money are directly related.

One way of measuring a person's demand for money is to measure his or her average holdings of money. To take a simple example, suppose that you receive a $900 monthly income that is deposited in your checking account on the first of every month, and that you spend it at a rate of $30 a day, every day, over a 30-day month. Halfway through the month, you will have spent half of your $900, so your bank balance will be $450. At the end of the month, it will be $0. The midmonth balance of $450 is your average balance. Now suppose that your income goes up by a third to $1,200, and your daily expenditure also increases by a third to $40. Your midmonth (and average) balance will then be $1200/2 = $600.

In this simple case, your average bank balance changes in the same proportion as your income—one third. But even if your daily pattern of receipts and payments were much more complicated, your average holding of money—your demand for money—and your income would still be directly related.

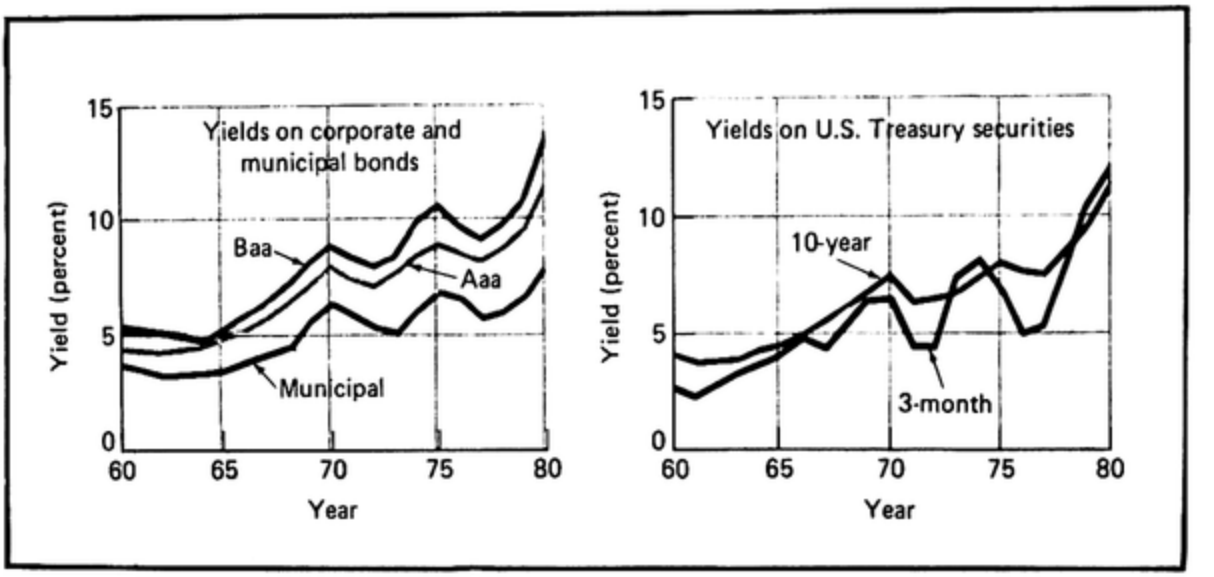
Income, however, is only one of two major influences on the demand for money. The other is the interest rate. To see why, remember the concept of opportunity cost. Holding money involves both a benefit and a cost. The *benefit* is convenience. Anyone who *holds* a stock of money can make deposits and pay bills quickly and conveniently. There is no need to buy and sell securities or transfer funds to and

Interest Rates and the Principle of Equal Advantage

Active investors, particularly financial intermediaries, are quick to move their funds around in search of higher rates of return or *yields*, as they are sometimes called. Their tendency to do this brings the yields on various kinds of securities into line with one another, after allowance for differences in risk, maturity, and other relevant considerations. You can see this in action in the two accompanying diagrams. The top one presents the returns on three different groups of long-term bonds over 20 years. Each series is an index number prepared by one of the well-known firms that provide information to investors. The top two series cover corporate bonds classified into two categories by Moody's investors' service: Baa (read "bee-double-aye") bonds, which are moderately high in risk, and Aaa (read "triple-aye"), which are quite low in risk. As you can see, the yields on the two kinds of bonds move together quite closely, but the riskier bonds offer a consistently higher yield to compensate their owners for bearing the risk. At the bottom of the diagram is a third series, covering municipal bonds classified as "high grade" by another investors' service, Standard and Poor's. This series fluctuates in the same pattern as the two corporate series, but is consistently below them. Does the differential in yields reflect greater security? Hardly. Even the most "high grade" municipalities may get into deep financial trouble someday, as New York and Cleveland did in the 1970s. In this case, the difference in yields reflects different tax treatment. Interest on municipal bonds is exempt from federal income tax. Because of this, the returns on municipal bonds are competed down until they offer the average bondholder the same after-tax return as taxable bonds. Of course, very wealthy people benefit the most from owning municipal bonds, since the tax rates they avoid are higher than those of the average bondholder.

The close relationship over time among the yields on various kinds of long-term securities reflects what economists call the *principle of equal advantage*. Differences in yield among assets reflect the value that the average investor places on security, tax treatment of interest, and other factors that distinguish one security from another. All yield differentials that do not reflect specific qualitative differences among kinds of instruments get competed away in the market as investors seek out the highest returns on their money. When the yields are in line with one another, no one can gain from shifting funds from one to the other. Assuming that the qualitative differences among assets remain unchanged, so do the differences among their yields. All rates of interest then move up and down together.

Another application of the principle of equal advantage explains the differing behavior of short- and long-term rates of interest. The bottom diagram shows the yields on two different kinds of U.S. Treasury securities. One of them consists of ten-year bonds, fairly representative of longer-term Treasury issues. The other consists of newly issued three-month Treasury bills. Both of these securities come from the same debtor, and



Illustrations of equal advantage

The yields on long-term bonds move together. Differences between them reflect risk and tax advantage. The yields on short- and long-term U.S. Treasury securities also move together. Since short-term securities are less prone to price fluctuations during the brief period they are held, they are less risky, and their yields are generally lower.

neither has any default risk, short of the risk that the entire social order will collapse.

There are three things to notice about these diagrams. First, short- and long-term interest rates generally rise and fall together, for the same reason that yields on Baa and Aaa bonds move together—equal advantage. Second, cyclical fluctuations in short-term rates are greater than those in long-term rates. Equal advantage explains this one, too. If you buy a short-term asset when its yield reaches a cyclical trough, you get this low return on your investment for only a short time. But if you buy a long-term asset at the low point of its yield, you are stuck with a lower than average return for a longer period. Thus, the long-term rate does not have to drop as far as the short to equalize the advantages. Third, the yield on short-term securities is usually below that on long. Their quick maturity minimizes the fluctuations in their prices during the time they are held. Most investors don't like capital losses and will accept a lower yield to avoid the risk of loss.

from a mutual fund when cash flows and ebbs. The *cost* of holding money is forgone interest. It is an opportunity cost. Most people's average checking account balances are roughly 20 percent of a month's income. Someone with a $2,000 monthly income holds an average balance of $400. If, instead, those funds were earning 5 percent in a savings account, they would provide an annual income of $20. This is the income forgone by holding money in a checking account. In this example, it

doesn't amount to much. But think of a large corporation with a monthly transaction flow of $1 billion. If it keeps an average balance equal to 20 percent of its transactions, it forgoes interest on $200 million. Since large amounts are involved, it can go directly to the securities markets to purchase Treasury bills (which come in $10,000 denominations). The return on bills is usually about twice that on small deposits in a savings institution. At 10 percent interest, the corporation's money holdings cost it $20 million a year. Because of the greater chances to earn interest on securities, large firms hold much smaller balances relative to transactions than households do. They keep their liquid assets in large certificates of deposits, Treasury bills, or other short-term assets. Some even buy securities late in the day and resell them in the morning to avoid holding large amounts of money overnight.

The cost of holding money, then, is the interest rate. When it changes, the demand for money will change. If the interest rate rises, the opportunity cost of holding money rises, and the demand for money will fall. If the interest rate falls, the opportunity cost of holding money falls, and the demand for money will rise. There is a *direct* relation between the demand for money and income, but there is an *inverse* relation between the demand for money and the interest rate.

Understanding how the demand for money is influenced by both the level of income and the interest rate can help you solve a puzzle. As GNP changes, the level of transactions and, therefore, the demand for money will change. For example, suppose that GNP rises. The level of transactions will rise, and the demand for money to finance them will also rise. If the supply remains unchanged, it would seem that the increase in the demand for money cannot be met. Does this mean the GNP cannot rise?

To answer that question, you need to understand the concept of ***velocity,*** or the rate of turnover ***of money.*** Velocity is the ratio of transactions to the quantity of money used in making them. It is determined by the receipts and payments patterns of firms, households, and governments. A given monthly or annual pattern of receipts and expenditures produces a given velocity or rate of turnover of the money supply. Think back to the earlier example of your average bank balance. Your transactions are measured by your monthly income of $900, while your average balance is $450. The monthly velocity of your money holdings is income/average bank balance = $900/$450 = 2. Your transactions are twice the level of your average bank balance. On an annual basis, the velocity is 2 × 12 = 24, since your annual income is 12 times your monthly income. Your money "turns over" for you twice a month or 24 times a year. You can finance a year's transactions of $900 per month × 12 months = $10,800 with a $450 average balance.

Following the same example, suppose that your income rises by one third to $1,200 per month and your average balance also rises by one third to $600. Your monthly income velocity is still 2 (1,200/600 = 2), and your annual velocity is still 24. With an unchanged pattern of receipts and expenditures and, therefore, unchanged velocity, you will not be able to finance your higher level of transactions without more money.

For the economy as a whole, the same kind of relationship holds between money velocity and GNP:

$$\text{Supply of money} \times \text{velocity} = \text{level of transactions or GNP.}$$

Going back to the original puzzle, how can GNP change if the supply of money is unchanged? The answer should be obvious: It can change only if velocity changes. If ve-

locity increases, with money turning over faster, then a given supply of money can finance a higher level of GNP. But why would velocity increase? Why, for example, might you reduce your average balance relative to your income? The answer is that you will do so if it is in your self-interest. Remember that holding money involves the opportunity cost of forgone interest. When the interest rate rises, individuals and firms have an incentive to cut back on their average balances. As individuals and firms economize on cash balances, average money holdings fall relative to transactions, and velocity increases.

This practice of cutting back on cash balances when interest rates are high has increased over the past two or three decades. This can be seen in the failure of the M1 money supply to keep pace with the growth in money GNP. It shows up clearly in Panels I and II of Figure 1, which compare the velocity of M1 to the rate of interest. Recall that M1 consists of currency and checking deposits. This kind of money yields little or no interest. Each cyclical upswing in interest rates has prompted both firms and individuals to find new ways to economize on such low-interest cash holdings. The velocity of M1 has increased enormously, paralleling the rise in interest rates. During the 1960s and early 1970s, there was some tendency for the rise in velocity to slow down when interest rates dropped, but this aspect of the relationship seems largely to have disappeared by the late 1970s.

A large part of the drop in demand for M1 money during the 1960s and 1970s was a switch from currency and checking deposits to savings and time deposits. Since these alternative deposits bear interest but are easily converted into spendable money, they are an attractive haven for temporary holdings. As you can see from Panel III in Figure 1, there was no uptrend in the velocity of M2 (which includes such interest-bearing deposits) during the 1960–1980 period. Moreover, there is an obvious tendency for the velocity of M2 to go down when the interest rate drops, and to go up when the interest rate rises. When interest rates are generally low, the convenience of interest-bearing bank deposits attracts and holds funds even though their yield is low. When interest rates are high, security yields typically rise relative to the return

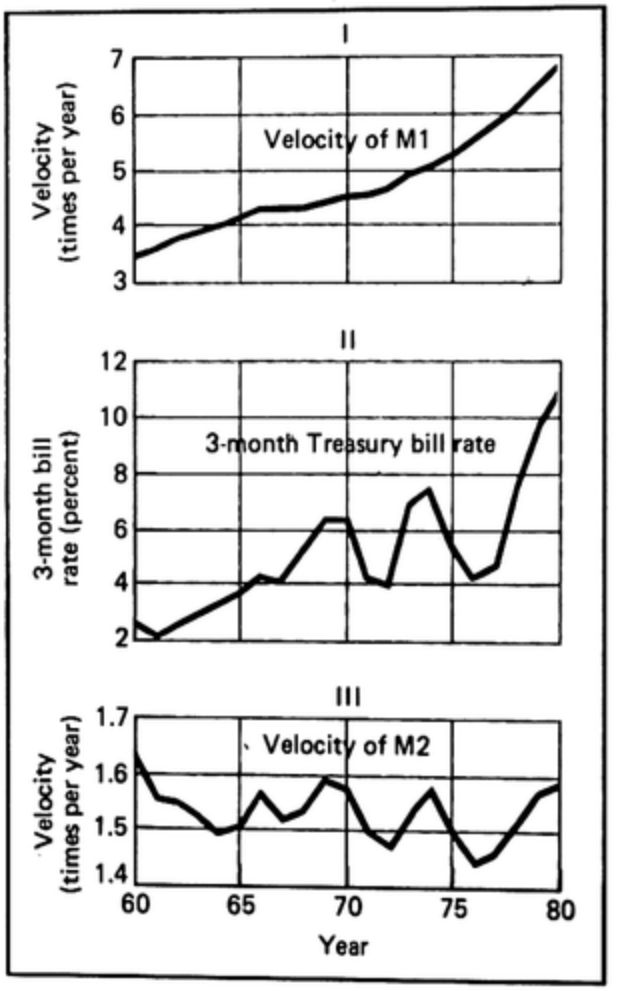


Figure 1 The velocity of money and the rate of interest

During the upsurge in interest rates in the 1960s and 1970s, individuals and firms learned to economize on M1, which bears little or no interest. Its velocity rose steadily, in a pattern that was not closely related to fluctuations in the rate of interest. The velocity of M2, much of which is interest bearing, showed no such uptrend, but it was closely correlated with cyclical fluctuations in the rate of interest.

Source: *Economic Report of the President.*

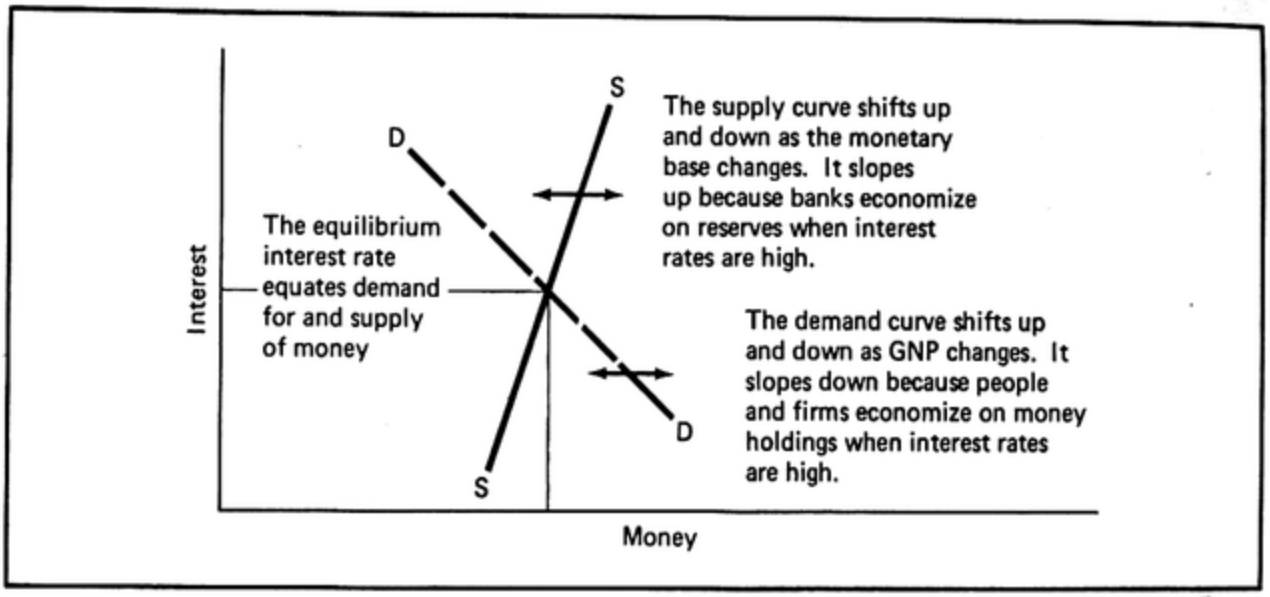


Figure 2 Determination of the equilibrium rate of interest

The equilibrium interest rate equates demand for and supply of money. The demand curve shifts in the same direction as GNP. The supply curve shifts in the same direction as the monetary base and responds to changes in other Federal Reserve policies. The higher GNP is, the higher the interest rate is. The higher the monetary base is, the lower the interest rate is.

on savings and time deposits, and people economize on their holdings of interest-bearing deposits as well as on currency and checking accounts.

The equilibrium rate of interest

You can now see the importance of the interest rate in determining velocity and, therefore, the demand for money. But what determines the interest rate itself? What sets its general level and makes it rise and fall?

The interest rate is a price—the opportunity cost of holding money. The equilibrium interest rate is determined by the supply of and demand for money. Both the supply and demand schedules are illustrated in Figure 2. Remember that the demand for money is influenced by the level of GNP and the interest rate. The *position* of the demand schedule depends on the level of GNP. A given demand schedule, such as *DD* in Figure 2, is drawn for a particular level of GNP. If GNP rises, then the demand for money will rise, and the demand schedule for money will shift to the right. If GNP falls, the demand schedule will shift to the left. The downward *slope* of the demand schedule reflects the inverse relationship between the interest rate and the demand for money. When the interest rate falls, the opportunity cost of holding money falls, and therefore the quantity of money demanded will rise. Given the level of GNP, then, the demand for money will be high when the interest rate is low, and low when the interest rate is high. A change in GNP will cause a shift in the money demand curve, but a change in the interest rate will cause a movement along it.

The supply schedule for money is labeled *SS* in Figure 2. Its *position* is determined by the size of the monetary base, the level of reserve requirements, and the interest rate at which depository institutions may borrow from the Fed. The supply schedule has an upward *slope*, because at high interest rates banks will hold fewer excess reserves, creating a larger amount of money from a given monetary base.

The equilibrium interest rate equates the demand for money to the supply. Knowing what influences the position of

the supply and demand schedules for money makes it fairly easy to see what causes the equilibrium interest rate to rise or fall. If GNP rises, the money demand schedule will shift right and the equilibrium interest rate will rise. If GNP falls, the money demand schedule will shift to the left, and the equilibrium interest rate will fall. If the Federal Reserve increases the monetary base, decreases reserve requirements, or lowers its lending rate, the money supply schedule will shift to the right, and the equilibrium interest will fall. If it lowers the monetary base, increases reserve requirements, or raises its lending rate, the money supply schedule will shift to the left, and the equilibrium interest rate will rise. These four cases are illustrated in Figure 3. Remember as you study these examples that when the equi-

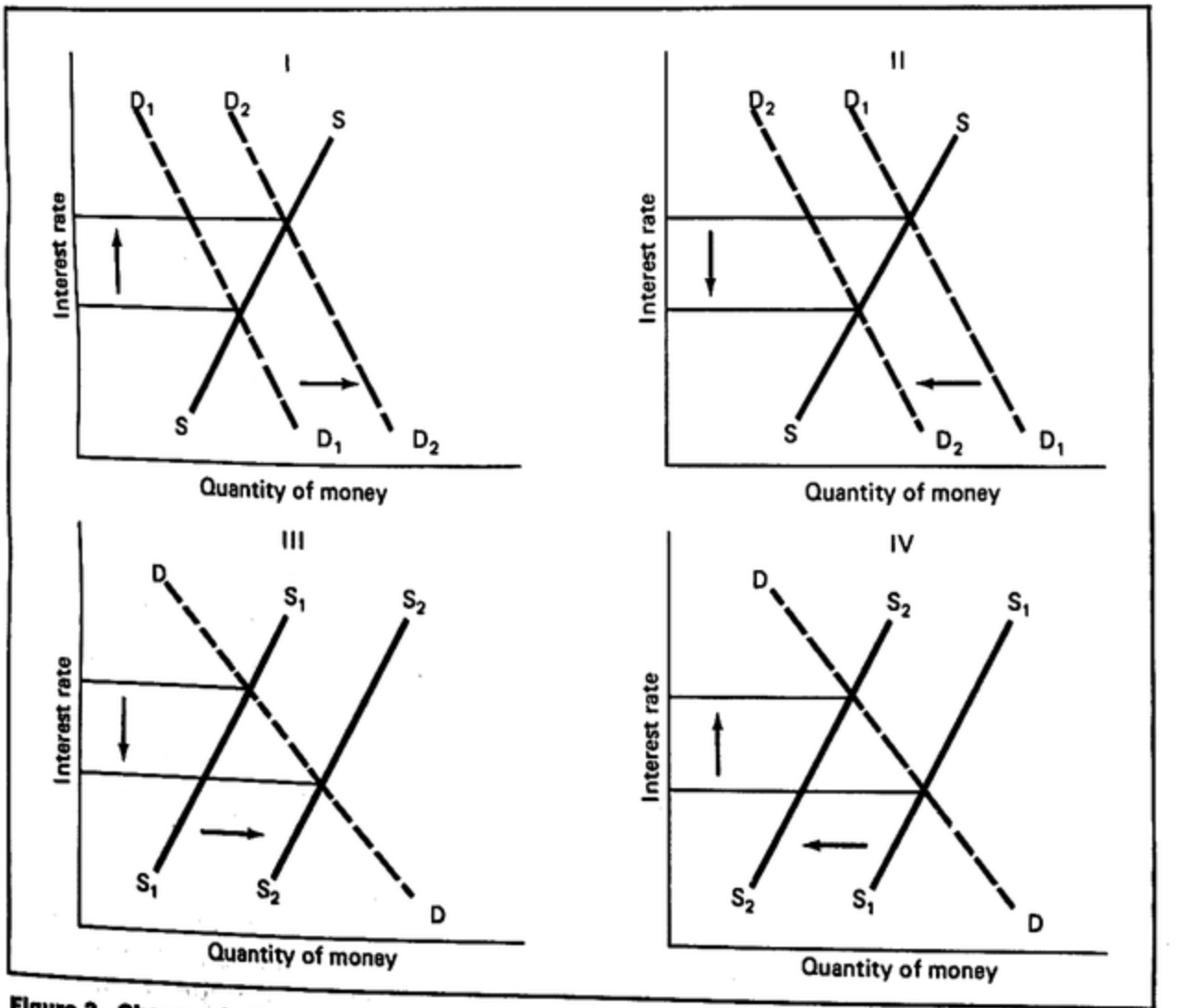


Figure 3 Changes in the equilibrium interest rate

When GNP increases, the money demand curve shifts to the right, and the equilibrium interest rate goes up, as in Panel I. When GNP drops, the money demand curve shifts to the left, and the equilibrium interest rate falls, as in Panel II. If the Federal Reserve increases the monetary base, reduces reserve requirements, or cuts its lending rate, the money supply curve shifts to the right, and the equilibrium interest rate falls, as in Panel III. If the Federal Reserve reduces the monetary base, raises reserve requirements, or increases its lending rate, the money supply curve shifts to the left, and the equilibrium interest rate rises, as in Panel IV.

librium interest rate changes, the velocity of money changes in the same direction.

The equilibrium rate of interest, then, is determined by the demand for and supply of money, and changes in demand and supply cause changes in the equilibrium interest rate. But this does not tell you what directly causes the interest rate to change. The next section takes a brief look at how interest rate changes come about.

Changes in the interest rate

Suppose that both the goods market and the money markets are in equilibrium. The economy is in multiplier territory, with sufficient unemployment to permit GNP to rise in response to an increase in demand. Now suppose that there is an upturn in business confidence, so that the demands for both fixed capital and inventories rise. The planned demand schedule for goods shifts up. This shift starts the multiplier process working, and GNP begins to rise toward a new and higher equilibrium. As these changes occur in the goods markets, other changes occur in the financial markets. During the expansion, when planned demand is greater than GNP, planned deficits are greater than planned surpluses. (If you don't remember why, look back to the chapter on the multiplier.) The demand for funds to finance deficit units is greater than the supply of funds from surplus units, and this puts pressure on the interest rate. As financial institutions attempt to ration the scarce supply of loanable funds, they raise interest rates.

The same process works when planned demand is declining. During such a contraction, planned deficits are smaller than planned surpluses, and the demand for loans is correspondingly smaller than the supply of loanable funds. Interest rates are cut by lenders, who attempt to attract borrowers.

The Federal Reserve and the rate of interest

As you know, changes in the demand for money are not the only kind of changes in the equilibrium rate of interest. Supply changes are equally important. If, for example, the Federal Reserve increases the monetary base or lowers reserve requirements, it provides excess reserves to the banking system and shifts the money supply schedule to the right. The banking system has excess lending capacity, and tries to increase its loans and holdings of negotiable debts. This has two effects. First, it lowers interest rates in all financial markets, as banks and other financial intermediaries compete with one another for business. Second, the additional bank lending increases the money supply through the process described in the last chapter. Similarly, when the Fed reduces the monetary base or increases reserve requirements, it forces the banking system to contract its loans and security holdings. This sends interest rates up and reduces the money supply.

As you learned from the last section, shifts in the planned demand schedule for GNP shift the demand for money and cause changes in interest rates. However, the Federal Reserve can keep interest rates unchanged by shifting the money supply schedule to offset the effects of the shifts in planned demand. Suppose, for example, that the planned demand schedule shifts upward. This tends to drive interest rates up because expansion in GNP creates excess demand for loans. But if the Fed simultaneously expands the monetary base, it permits the banking system to supply the loan demand without a rise in the interest rate. In fact, the equilibrium interest rate may remain unchanged if the *SS* schedule in Figure 2 is shifted to the right to keep pace with the *DD* schedule.

This also works in reverse. If there is a downward shift in the planned demand schedule, planned deficits will drop relative to planned surpluses. There will be an

excess supply of funds to lend, and the interest rate will fall. If the Fed contracts the monetary base, it can curb this drop in interest rates by reducing the lending capacity of the banking system and shifting the money supply schedule to the left.

The reserve position of the commercial banking system is an excellent barometer of the demand-supply balance in the credit markets and of pressures on the interest rate. When loan demand is low relative to supply, many banks hold excess reserves. When demand rises, banks increase their lending. Excess reserves fall off as required reserves go up, along with the money supply. Some banks are even forced to borrow reserves from the Fed to meet their reserve requirements.

The best summary measure of the banking system's ability to supply new loan demand is a quantity called *net free reserves*. This is simply the difference between the banking system's excess reserves and its borrowing from the Fed. When it is positive, the average bank has excess reserves; when it is negative, the average bank is in debt to the Fed. Figure 4 shows the fluctuations in net free reserves over a 20-year period. The top part of the diagram shows the yield on Treasury bills for comparison. Remember that when net free reserves are positive, the banking system has excess loan capacity. Such periods are times of "easy money." As you can see, they coincide with, or lead slightly, the periods of low interest rates. When net free reserves are negative, the banking system is so "loaned up" that some banks are borrowing reserves. These dips in net free reserves match up with cyclical peaks in interest rates. Thus, the fluctuations in interest rates seem to reflect faithfully the supply-demand situation in the loan market.

A moment's reflection will convince you that the Federal Reserve must therefore largely control the interest rate, at least in the short run. In the last chapter,

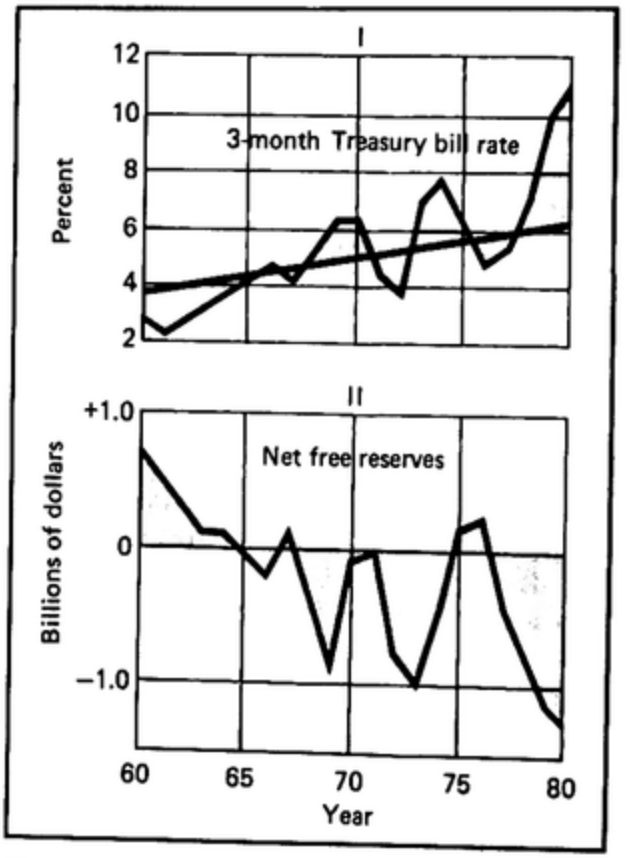


Figure 4 Bank reserves and interest rates

Fluctuations in short-term interest rates follow closely the net free reserves that the Fed supplies to the banking system. Plentiful free reserves lead to low interest rates, and scarce reserves lead to high interest rates.

Source: *Economic Report of the President.*

you learned how Federal Reserve purchases of government securities create excess reserves for the banking system. Fed sales of securities reduce bank reserves. It follows that the Fed can make net free reserves positive by buying a sufficient volume of securities, and negative by selling enough securities. By changing the monetary base in this way, it can affect the availability of credit and the direction of movement of interest rates. Changes in reserve requirements and the Fed's lending rate have similar effects.

Think about this first in terms of a cyclical upswing. Suppose that the Fed provides reserves, so that the banks can fi-

nance the excess loan demand with newly created money. Then the rate of interest need not rise during expansion. During a downswing, the Fed can keep the banking system from developing excess loan capacity if it reduces reserves by selling government securities. This will keep the interest rate from dropping. If it chooses, the Fed can iron out much of the cyclical fluctuation in the interest rate.

In fact, the Federal Reserve doesn't usually operate this way. It rarely creates reserves fast enough to keep interest rates from rising during an expansion, and it never tries to keep interest rates from falling during a contraction. The reason is that a rising interest rate slows down the process of expansion and a falling interest rate slows down the process of contraction. By letting interest rates rise and fall as income rises and falls, the Fed hopes to make GNP fluctuations less severe than they would otherwise be.

Interest and expenditures

In the first main section, you learned how changes in GNP lead to changes in the interest rate. This section explores how changes in the interest rate also affect GNP.

Consumer demand

Remember that the share of disposable income that households consume fluctuates around an average value of about 93 percent. The ordinary range of variation on either side of that average is a couple of percentage points. You might suppose that some of this fluctuation in the share of income consumed is correlated with changes in the interest rate. After all, a higher interest rate raises the cost of consumer credit and the returns from saving. Therefore, people might be expected to spend less when the interest rate is high, and more when it is low.

In fact, this does not seem to be the case, at least not in any obvious way. This may surprise you if you have ever heard anyone complain about high mortgage interest rates. What about the building industry? Isn't housing demand always depressed when credit is tight?

The answer is yes, of course. Mortgage rates are the most important determinants of the demand for housing. But home buying is the one household activity that is classified as investment, not consumption. The *use* of the house is consumption, but the *purchase* is an investment.

Residential investment and the interest rate

The accompanying box shows the figures needed to get the monthly payment on a mortgage or installment loan in which principal and interest are paid in a series of equal amounts over the life of the loan. As you can see, a $50,000, 30-year mortgage, which might finance a modest suburban house, would cost the borrower $438.50 a month over the life of the loan at 10 percent interest. (This does not count taxes and insurance, or the many fees paid when the loan is taken out.)

Of particular importance for the interest sensitivity of expenditure is how the monthly payment changes when the interest rate changes. Look at the 30-year column. As the box explains, when the interest rate goes from 10 to 15 percent, the monthly mortgage payment goes from $438.50 to $632.00, a 44 percent increase. On a five-year loan of $50,000, the monthly payment only rises by 12 percent, from $1,059.00 to $1,184.50.

The greater interest sensitivity of longer-term loan payments explains why people are more influenced by interest rates when they are thinking about borrowing to buy a house than to buy a car.

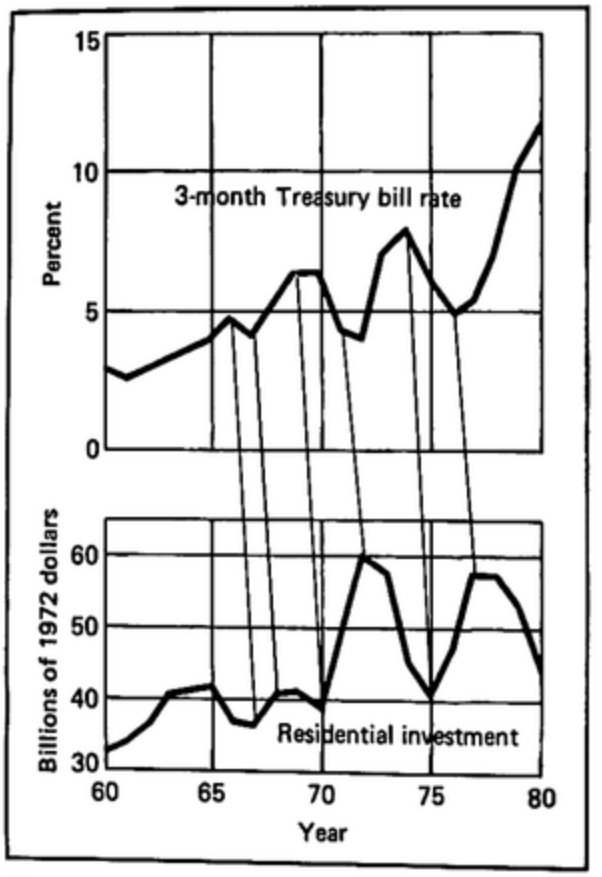


Figure 5 Interest rates and residential investment 1960–1980

The major peaks and troughs in the rate of interest are followed by fluctuations in residential investment with a one-year lag. High interest rates depress home building and low interest rates encourage it.

Source: *Economic Report of the President.*

Most car loans have a term of three to five years. Home mortgages usually run 20 or 30 years to maturity.

Since consumer demand for housing is closely tied to interest fluctuations, so is new construction. Housing isn't built unless it seems likely to be bought before or soon after it is finished. The home building industry operates with a lot of borrowed funds and little equity. If the builder doesn't sell almost immediately, carrying costs wipe out all the profit.

Figure 5 shows the fluctuations in residential investment in constant dollars from 1960 to 1980. This is new housing constructed, and amounts to about 25 percent of total fixed investment in a prosperous year for the building industry. The top part of the figure reproduces the Treasury bill rate as an indicator of credit tightness. As you can see, construction of new housing fluctuates in the opposite direction from the interest rate, with about a one-year lag. This interest responsiveness of new home building is the most important link between the credit markets and the demand for goods.

Business Investment

The other major link between monetary conditions and the market for goods and services is the interest sensitivity of business fixed investment, that is, investment in plant and equipment. This is less easy to see than the housing link because the demand for business investment is in a sense more complicated, more dependent on a variety of influences, than is the demand for housing. *How much* housing is needed depends on such factors as population and wealth, which change only slowly over time. Within the limits set by population and wealth, the *timing* of housing construction is determined mainly by credit conditions. Business investment is different. Manufacturing in particular undergoes large cyclical swings in the demand for its output. The question is not just *when* to expand, but *whether* there is a need to expand at all. Unlike housing investment, then, business investment patterns are dominated both by fluctuating demand for output and by credit conditions.

Most theories of business investment that attempt to sort out the various influences go something like this. Firms project future demand or market conditions and then calculate the likely rate of return on various expansion and modernization projects. Those projects that offer a return greater than the cost of credit are undertaken. Those that do not are screened out. The degree of use of existing capacity is

Monthly Payments of Principal and Interest per $1,000 Borrowed

	5-Year Loan	15-Year Loan	30-Year Loan
10 percent interest rate	$21.18	$10.73	$ 8.77
15 percent interest rate	$23.69	$13.97	$12.64

These figures are calculations of the monthly payments of principal and interest per $1,000 borrowed, for loans of varying length and various interest rates. Using them, you can compare how a change in interest rates will affect loans of varying maturities.

Consider first a 30-year loan of $50,000. At 10 percent interest, the monthly payment would be:

$$50 \times \$8.77 = \$438.50.$$

If the interest rate were instead 15 percent, the monthly payment would be:

$$50 \times \$12.64 = \$632.00.$$

The percentage increase is:

$$193.50/438.50 = 44 \text{ percent.}$$

Now consider a 5-year loan of $50,000. At 10 percent interest, the monthly payment would be:

$$50 \times \$21.18 = \$1,059.00.$$

If the interest rate were instead 15 percent, the monthly payment would be:

$$50 \times \$23.69 = \$1,184.50.$$

The increase in monthly payments is only 12 percent.

The 5 percentage point increase in interest has a much larger impact on the 30-year loan than on the 5-year loan because the major part of the payment on a long-term loan is interest, while on a short-term loan it is principal.

one key variable in determining the expected return of return. When capacity is underused, it is not profitable to undertake large-scale expansion, even if interest rates are low. Why build more capacity when the firm cannot use the capacity it already has? When capacity is stretched to the limit, expansion promises a higher return. In this case, expansion may be profitable even if interest rates are high and credit is therefore costly.

Because of this interaction among the various influences on business investment, the influence of changes in credit conditions must be looked at indirectly. Economic statisticians try to do this by deriving equations to predict the response of investment to changes in capacity utilization, profitability, and other variables, including interest rates. These equations clearly indicate that the interest rate matters. However, they give differing measures of the importance of interest rates, depending on the precise way in which expectations and other factors are taken into account. This creates a range of uncertainty about the precise quantitative impact of monetary developments on business investment. It is clear, however, that there is an impact.

Combining the markets

We have now reached the climax of a long and intricate tale, in which all the threads of the plot come together for resolution. The principal subplots are found in the chapters on equilibrium, the multiplier, inflation, and the financial markets. The actors are the participants in the financial markets (including the Federal Reserve) and the participants in the goods markets (including the federal government).

The monetary feedback

It is clear from the preceding sections of this chapter that the link between the financial and goods markets is a two-way street. It is illustrated in Figure 6. The equilibrium rate of interest is determined at the intersection of the money demand and supply schedules. Equilibrium GNP is determined at the intersection of the planned demand schedule and the line of equality. These determination processes are represented by the circled diagrams at the top and bottom of Figure 6.

The financial and goods markets are linked by two distinct paths. One, which we will call the ***interest effect,*** runs from the interest rate to GNP by way of the planned demand schedule. The interest effect is negative. A rise in the interest rate shifts the planned demand schedule down and lowers equilibrium GNP. The other, which we will call the ***GNP effect,*** runs from GNP to the rate of interest by way of the money demand curve. The GNP effect is positive. A rise in GNP shifts the demand curve for money to the right and raises the rate of interest.

Together, the changes that make up the interest effect and the GNP effect form a closed path, from interest to GNP and back to interest. This closed path is a lot like the "feedback loops" that engineers use to stabilize mechanical and electronic

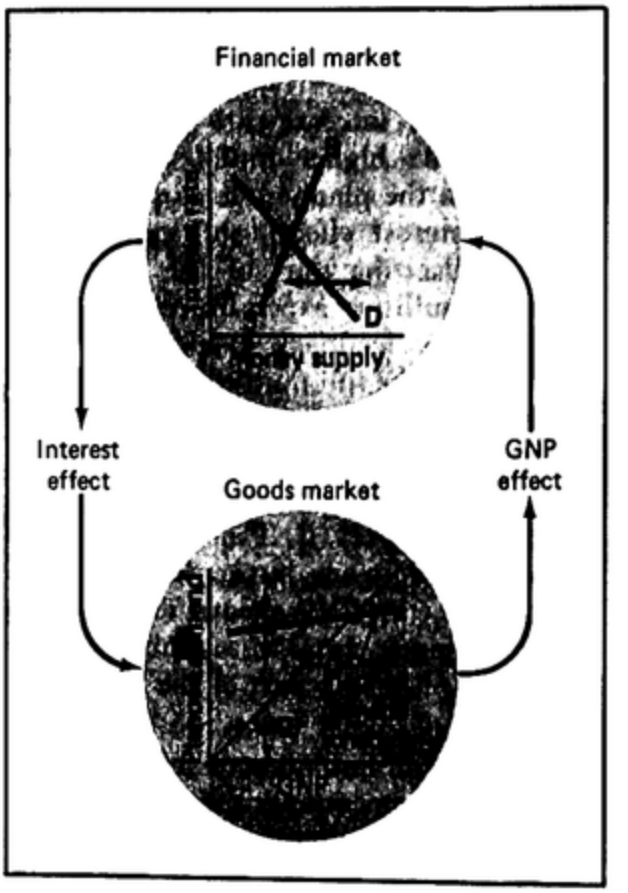


Figure 6 Simultaneous determination of GNP and the interest rate

The equilibrium interest rate is determined at the intersection of the money demand and supply curves, GNP at the intersection of the planned demand schedule and the line of equality. But the two processes interact, since changes in the interest rate shift the planned demand schedule, and changes in GNP shift the money demand curve. We will call the impact of interest on GNP the *interest effect*, and the impact of GNP on the interest rate the *GNP effect*. The two effects together make up the *monetary feedback*.

equipment. It is thus often called the ***monetary feedback*** and is a source of stability in the economy.

To see why it is a stabilizing force, consider the following examples. In each case, assume that the economy is initially in the territory of the multiplier process, so that shifts in planned demand affect real GNP.

1. There is an autonomous rise in the planned demand schedule because of a rise in military expenditures. This directly raises GNP, but it indirectly

raises the interest rate (the GNP effect). The higher interest rate feeds back on the planned demand schedule (the interest effect), shifting it down and offsetting part of the impact of higher military expenditures. Thus, the monetary feedback keeps GNP from rising by the full multiplier response to the autonomous spending change.

2. There is an autonomous drop in planned demand coming from an increase in taxes. GNP declines as a direct consequence, but so does the interest rate (the GNP effect). This drop in the interest rate raises planned demand, offsetting part of the multiplier impact of higher taxes.
3. There is an autonomous leftward shift in the money supply schedule because of Federal Reserve open market policy. This directly raises the interest rate. Because of the interest effect, GNP drops. Because of the GNP effect, the money demand schedule shifts to the left, lowering the interest rate and offsetting part of the direct impact of the shift in the money supply curve.
4. There is an autonomous leftward shift in the money demand curve because of increased use of credit cards. This directly lowers the interest rate. Through the interest effect, this raises GNP. But the GNP effect shifts the money demand curve back to the right, offsetting part of the autonomous impact on the interest rate.

Notice that each of these examples incorporates an autonomous change that directly affects either equilibrium GNP or the equilibrium interest rate. Because of the monetary feedback, however, there are also indirect effects on both GNP and the interest rate. The indirect effects always offset part of the direct impact and stabilize the economy. Note also that whether the autonomous change takes place in the financial market or in the market for goods and services, it must eventually affect both markets because of their interdependence.

This discussion is, of course, greatly simplified and only compares equilibrium positions. The next section gives one illustration of the process of moving from one equilibrium to another.

The Federal Reserve and GNP

Understanding the connections between the financial and goods markets makes it easier to see exactly what the Federal Reserve is doing when it conducts open market operations, sets reserve requirements, and chooses a discount rate. Many of its day-to-day policies are aimed at stabilizing the financial markets themselves—responding to sudden changes in the demand for and supply of credit to prevent unwanted fluctuations in security prices. But ultimately, its eye is on GNP. It hopes to stabilize the market for goods by operating in the financial market. It is thus conducting ***monetary policy,*** one of the two main branches of stabilization policy.

Suppose, for instance, that the Board of Governors of the Fed wishes to reduce GNP because it believes that the economy is dangerously close to the boundary of inflationary territory. How may it do this?

There are three things it could do: instruct the Federal Open Market Committee to sell securities to the public, raise reserve requirements, or raise the discount rate at which it lends reserves to the banking system. The first of these would directly reduce the actual reserves of the banking system and, therefore, the size of the monetary base. The second would increase required reserves. The third would make it more costly for banks to borrow reserves from the Fed itself.

Any of the three would shift the money supply curve to the left. Ultimately, this shift would lead to a contraction in GNP.

To see why, think in terms of a concrete example, an open market sale of government securities by the Fed. This starts off a sequence of events in the financial and goods markets:

1. The Federal Reserve sells securities to the public and thereby reduces the reserves of the banking system.
2. The monetary base declines, the money supply curve shifts to the left, and the equilibrium interest rate goes up.
3. Banks are forced to reduce their outstanding loans and sell securities to try to replenish their reserves.
4. This reduces the money supply and, therefore, required reserves, bringing them into line with the lower actual reserves. It also creates a shortage of loan funds.
5. Would-be borrowers compete for the limited supply of credit and drive up the market interest rate.
6. Higher interest rates, with some lag, lead to a reduction in residential and business fixed investment.
7. This drop in planned demand sets off a multiplier contraction, as the Fed intended all along.
8. Because of the interest and GNP effects, there is a series of secondary consequences for both the interest rate and GNP.
9. In the end, the interest rate is higher, and the GNP lower.

This process, which begins with an open market sale and ends with a drop in GNP is illustrated, in Figure 7. It shows how indirect the path is by which the Federal Reserve seeks to control GNP. A similar sequence of events could be illustrated for a rise in the reserve requirement, which also forces a loan contraction; or for a rise in the discount rate, which reduces the banks' willingness to hold reserves borrowed from the Fed. Of course, a parallel series of arguments could be developed to show why an open market purchase of government securities, a reduction in reserve requirements, or a reduction in the discount rate could be expected to lead to a multiplier expansion.

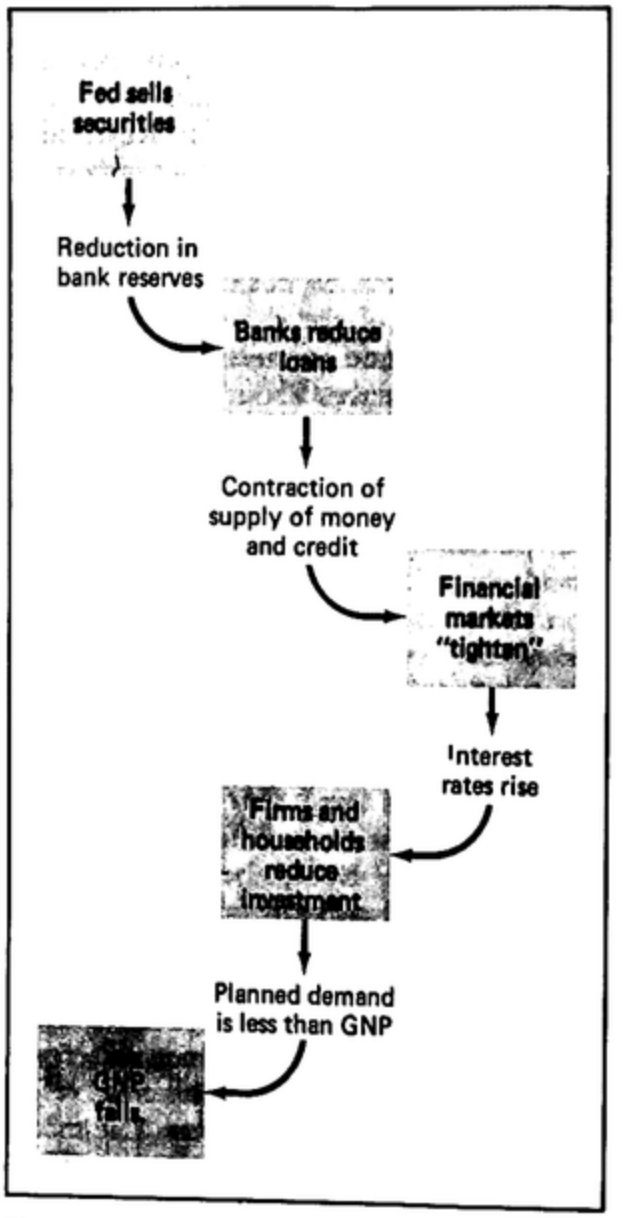


Figure 7 The impact of a Federal Reserve sale of securities to the public

By selling government securities to the public, the Fed sets off a chain of events that produces a reduction in GNP.

Keynesianism and monetarism

Understanding the connections between the financial and goods markets is also helpful for understanding the Keynesian-monetarist debate, the most important macroeconomic controversy of recent decades. There are many subtleties and shades of opinion on both sides of this issue, but most of the differences between ***Keynesianism*** and ***monetarism*** can be reduced to differences about the strength of the interest and GNP effects—the monetary feedback.

The main structure of the controversy will be easier to follow if you first understand some of its building blocks. Turn your attention to Figure 8. Panel I illustrates the two groups' differing views about what goes on in the financial markets. As you can see, the monetarists think the money demand curve is quite steep, the Keynesians think that it is less steep. Essentially, this is a difference in views about velocity. Remember that a given money demand curve is drawn for a given level of GNP. If the money demand curve is very steep, then the demand for money at a given GNP doesn't change much when the interest rate changes. Since velocity is the ratio of GNP to money demand, velocity doesn't change much either. The monetarists think that velocity is quite stable—unresponsive to changes in the interest rate. The Keynesians, by contrast, think that it moves readily in the same direction as the interest rate.

This difference about velocity has some important implications. The first can be seen in Panel I of Figure 8. For the Keynesians, a shift in the money supply curve won't lower the interest rate very much. A small drop in the interest rate will make people want to hold a lot more money, causing velocity to drop sharply. For the monetarists, the same shift in the money supply schedule will lower the interest rate a lot. Since velocity is so unresponsive to the interest rate, it takes a large interest rate drop to make people hold much additional money at an unchanged volume of transactions. You can see immediately from this that monetarists accord the Fed a much greater degree of short-run control over the interest rate than the Keynesians do. According to the monetarists, relatively small changes in the money supply produce relatively large changes in interest. According to the Keynesians, relatively large changes in the money supply produce relatively small changes in interest.

The second consequence of the differing views about the demand for money can be seen in Panel II of Figure 8. If the demand curve is very steep, the interest rate goes up a lot for a given rightward shift in the curve. This is the monetarist view. If it is flatter, the interest rate response to the same demand shift is smaller. This is the Keynesian view. Obviously, these opposing views reflect a difference in belief about the strength of the *GNP effect*. The monetarists think it is strong, the Keynesians believe it is weak.

Finally, look at Panel III of Figure 8. It illustrates another major difference in belief between the two groups. The Keynesians think that the planned demand schedule shifts relatively little when the interest rate changes a given amount. The monetarists think it shifts a lot for the same change in the interest rate. This, of course, is a difference about the strength of the *interest effect*.

These two major differences collapse into one. Both concern the strength of the *monetary feedback*. The monetarists think that both its links are very strong; the Keynesians believe that they are less strong. It is basically an empirical question, a question of degree. But it is difficult to resolve because the overall process that determines GNP and the interest rate is both complex and subject to autonomous

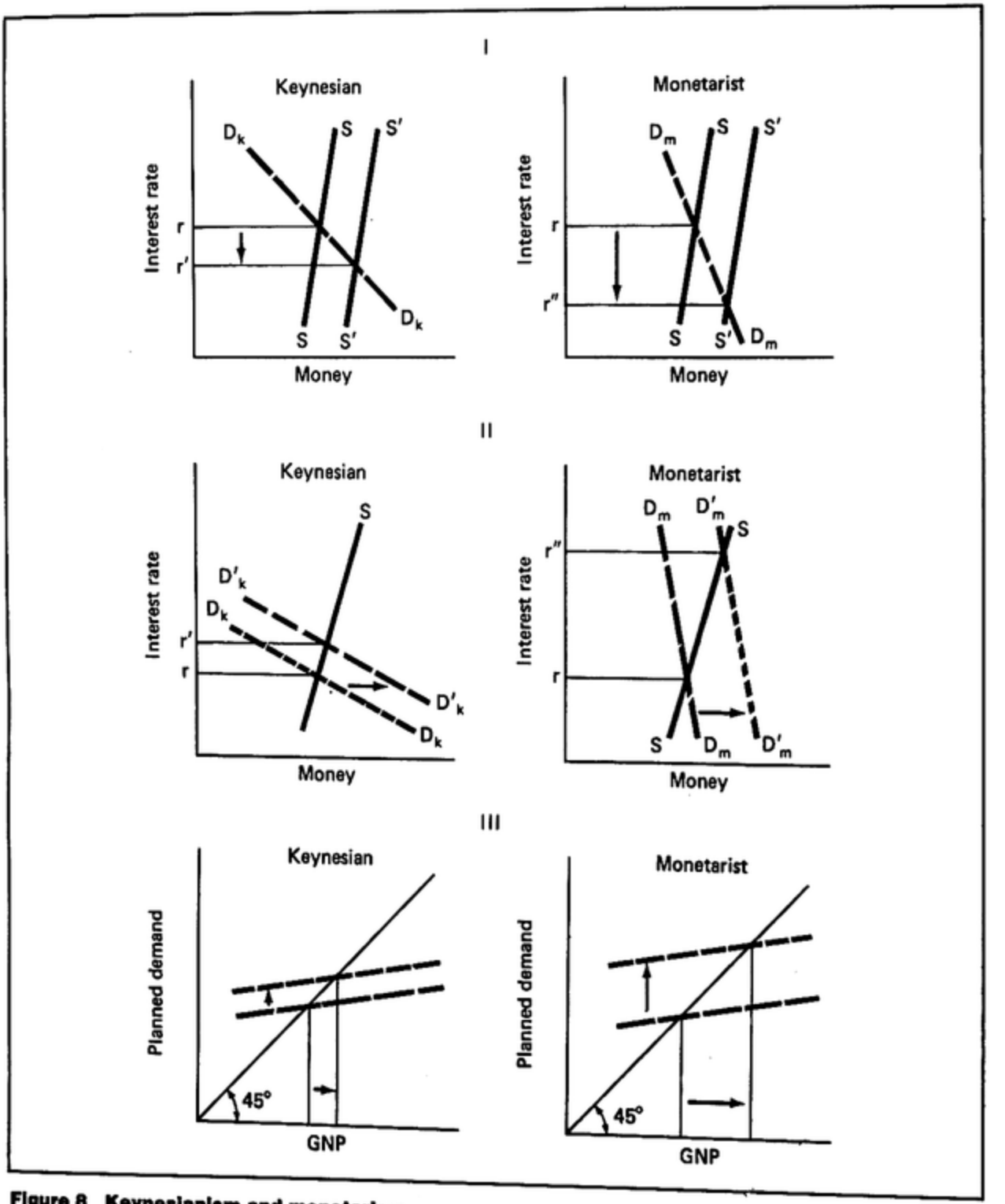


Figure 8 Keynesianism and monetarism

The monetarists think that the demand curve for money is relatively steep (D_m-D_m) rather than relatively flat (like the Keynesian demand curve D_k-D_k) (Panel I). This makes them expect a bigger rather than a smaller change in the interest rate when the money supply changes—from r to r'' rather than from r to r'.

Another implication of the different beliefs about the steepness of the money demand curve comes out when it shifts (Panel II). The relatively flat Keynesian curve produces a relatively small rise in the interest rate (r to r'). The steeper monetarist curve produces a larger change (r to r'').

The Keynesians and monetarists also differ about the strength of the interest effect (Panel III). For a given drop in the interest rate, the Keynesians expect a relatively small upward shift in the planned demand schedule and, therefore, a moderate increase in GNP. For the same drop, the monetarists expect a much larger shift in the planned demand schedule and a consequently larger increase in GNP.

influences that are hard to identify and measure.

Because it is an empirical question, don't think it is unimportant. The major controversies about both the source and the control of economic instability revolve around it.

Suppose, for example, that the monetary base fluctuates erratically because the Board of Governors of the Fed continually changes its mind about whether the economy ought to have a larger or a smaller money supply. Will this indecisive behavior produce corresponding fluctuations in GNP? The answer to this question depends on whether you ask a Keynesian or a monetarist.

The Keynesian will say no for the following reasons: Remember that velocity is quite flexible. At a given level of GNP, shifts in the money supply curve will be absorbed by changes in money demand without much change in the interest rate. Moreover, since the interest effect is weak, these interest fluctuations will have little impact on the goods market. Therefore, the results of instability in the money supply curve will be confined mainly to the financial markets, and will not even cause very large changes in the interest rate at that. Their impact on the goods market will be minimal.

Nonsense, the monetarist will respond. A careful look at the history of velocity will show that it is almost independent of the interest rate. Instability in the money supply curve will therefore produce large swings in the interest rate. And a similarly careful look at the history of interest rates and the demand for goods will show that the interest effect is very large. Shifts in the money supply curve will therefore produce large fluctuations in GNP. Instability in Federal Reserve policy means instability in both the financial and goods markets.

Suppose you also ask what will happen if the planned demand schedule fluctuates erratically and autonomously—say, because defense expenditures go up and down from year to year. Will this cause instability in both the goods and the financial markets?

The Keynesian answers yes. Autonomous demand changes will be amplified by the multiplier and produce amplified changes in GNP. Of course, these ups and downs in GNP will be reflected in the demand for money. But velocity is not a natural constant. It is really very flexible, so that changes in the demand for money can be contained by relatively small reactions in the interest rate. And since the interest effect is weak, the feedback to the goods markets will also be weak. The autonomous demand fluctuations will therefore cause major instability in the goods markets and some instability in the financial markets. The monetary feedback will definitely not stabilize GNP because it is too weak.

The monetarist will again explain the monetary feedback, less patiently this time, and argue that it will contain the disturbance. Because of the GNP effect, the initial disturbances in the goods market will be matched in the financial market. Because of the interest effect, the financial reaction will feed back on the goods market and stabilize it. The fluctuations in defense expenditures will simply *crowd out* private demand. Only the *composition* of demand will be altered. Remember that if military purchases go up and planned demand rises, so will the interest rate. The inflexibility of the velocity of money will see to this. A higher interest rate will reduce private investment demand, making room for the military purchases, and leaving GNP relatively unchanged. The ***crowding out*** may not be complete, but it will be substantial, and the multiplier will be largely

nullified by the monetary feedback. Autonomous changes in demand will shift the planned demand schedule all right, but the financial market's reaction will shift it back.

You probably cannot find a real Keynesian and a monetarist who will take precisely these positions. They represent poles of an argument that also has a middle ground. But the debate is often heated, and rightly so. Ultimately, it is a central issue for public policy, since the relative impacts of monetary and budgetary changes is a major question that must be faced when the government tries to stabilize the economy. We shall return to this issue in the next few chapters.

Money and Inflation

Toward the end of the chapter on inflation, a major issue was left hanging—the relationship between money and the inflationary process. Now that you understand the connections between the financial and goods markets, it is possible to resolve it.

Until now, no explicit mention of prices has appeared in this discussion of money, interest, and GNP. When looking at the inflationary process, however, we can hardly ignore prices. To integrate the level of prices and their rate of change into the analysis of goods and financial markets, we must discuss two new concepts, the ***real stock of money*** and the ***real rate of interest.***

Economists often find it important to distinguish "real" magnitudes, expressed in dollars of constant purchasing power, from "nominal" or "money" magnitudes, expressed in terms of dollars whose value varies inversely with the price level. You are already familiar with the distinction between real GNP and nominal, or money, GNP. The difference between real and money wages similarly depends on whether the dollar yardstick used to measure them is constant or variable in its ability to buy goods. Changes in real magnitudes have been adjusted for price changes. Changes in nominal or money magnitudes have not. With this in mind, it is not hard to understand what the real stock of money is.

The ability of any given nominal money supply to finance transactions depends on its command over goods and, therefore, on the price level. A given stock of money (whether M1, M2, or some other measure) is large relative to real GNP if the price level is low, but small if it is high. This causes ambiguity in the concept of velocity. The ambiguity disappears if you always think of velocity as the ratio of *real GNP* to the *real stock of money*, defined as the *nominal* or actual amount of money divided by the GNP deflator. The real money supply is then measured in dollars of constant purchasing power. The nominal stock is measured in dollars of changing purchasing power.

The real rate of interest is a little more complicated. Suppose that you are in debt and pay 15 percent interest on your borrowed principal. Suppose also that the annual rate of inflation is 10 percent, and has been for a long time, so that the 10 percent figure is firmly fixed in your expectations. The purchasing power of the principal you owe is falling at a current rate of 10 percent a year. Because of this decline in the real amount of your indebtedness, it only costs you 5 percent a year to stay in debt. The *real* rate of interest is therefore only 5 percent, even though the *nominal* rate is 15 percent. The real rate is the nominal rate minus the expected rate of inflation.

Now think about money and inflation. Suppose that the expected rate of inflation is zero, and that the financial markets and multiplier process are in equilibrium, but that the economy is in the territory of the inflationary process. People don't expect

prices to rise, but they do anyhow. Demand for GNP is high enough to keep the unemployment rate low, money wages are rising faster than productivity, and costs are pushing prices up.

Can this process continue for long? That all depends on what happens to the nominal money supply. Suppose first that the Fed keeps the monetary base, reserve requirements, and the discount rate unchanged. Then the supply curve for *nominal* money will be fixed. But think what is happening to the supply curve for *real* money. Since the purchasing power of the nominal money stock is falling, the real money stock must be falling too. Therefore, the supply curve of real money must be shifting steadily to the left as the price level rises. The drop in money supply relative to GNP will force the nominal interest rate up. As long as the inflation remains unexpected, a higher nominal rate of interest means a higher real rate of interest.

The interest effect will transmit these changes in the financial market to the goods market. As the real interest rate rises, the planned demand schedule will shift down. Eventually, the economy will move out of the territory of the inflationary process. Therefore, if the Fed does not allow the supply curve of nominal money to keep pace with inflation, the resulting *destruction of real money* will stop the inflation sooner or later.

Suppose, however, that the Fed increases the monetary base to keep pace with inflation. If it does, the supply curve of real money will be unaffected by the inflation, and there will be no money supply-demand imbalance to force interest rates up and stop the inflation.

If the Fed persists in such a policy of accommodating inflation, eventually the inflation will come to be expected. When borrowers and lenders both come to expect inflation, nominal rates of interest will move up to keep pace with the expected rate of price increase. But the higher nominal rate will have *no interest effect* on the goods market because *the real rate of interest will be unchanged.* This is the rate relevant for investment decisions.

If the Phillips curve shifts to keep up with expectations, the inflation will get worse. As the Fed persists in its policy of maintaining an unchanged real monetary base, it will have to keep increasing the nominal monetary base along with inflation. As long as the unemployment rate remains below the *natural rate* (the rate at which actual and expected rates of inflation are equal), the actual and expected rates of inflation will get larger and larger. So will the nominal rate of interest. Anyone who measures the tightness of money by looking at the nominal rate of interest will think that the Fed is pursuing an active anti-inflationary policy. In fact, however, it is *actively promoting inflation* by keeping the real rate of interest from rising. If it were to stop increasing the monetary base, it would bring the inflation to an eventual end by sufficiently lowering the ratio of real money to GNP.

In one sense, every persistent inflation is a monetary inflation. Money incomes rise automatically when prices rise. But the nominal money supply does not. If the nominal money supply keeps up with inflation, it is because the monetary authorities are allowing this to happen. In our economy, it is because the Federal Reserve is raising the monetary base as prices go up. During the persistent inflation of the 1970s, the Fed was criticized by monetary economists for doing just this. By continually expanding the nominal money supply, it systematically thwarted the tendency for inflation to bring itself to an end.

Money and interest in the long run

Paradoxically, one of the central beliefs of monetarism is that in the long run, the money supply determines only the price level and the level of money wages. This is agreed to by nearly every economist who thinks that the market system has a built-in tendency toward full employment. Real GNP, the real interest rate, the real wage rate, and all of the other real variables (including the real money supply) are independent of the nominal money supply. This proposition is known as the ***neutrality of money,*** since the nominal money supply has a neutral role in determining real economic variables.

Full employment is, of course, a slippery concept. A zero rate of unemployment is impossible, and there is no rate that everyone will agree is *the* full-employment level. One approach is to define full employment as the natural rate of unemployment, at which the expected and actual rates of inflation are equal, and the rate of inflation has no tendency to rise or fall.

Corresponding to the full-employment rate would be the full-employment level of GNP. If full employment is to be achieved, the planned demand schedule must cut the line of equality at the level of income corresponding to full-employment GNP, as shown by PD_1, in Figure 9. For this full-employment level of GNP to be reached, however, several other conditions must hold. For example, since the position of the planned demand schedule is influenced by the interest rate, the only possible interest rate in the long run is the one associated with the full-employment planned demand schedule. The money demand schedule must also correspond to full-employment GNP. Finally, the real money supply schedule must cut the money demand schedule at the point necessary to achieve the appropriate interest rate. It seems improbable that all of these conditions—the correct planned demand schedule, interest rate, and money supply and demand schedules—will come neatly together.

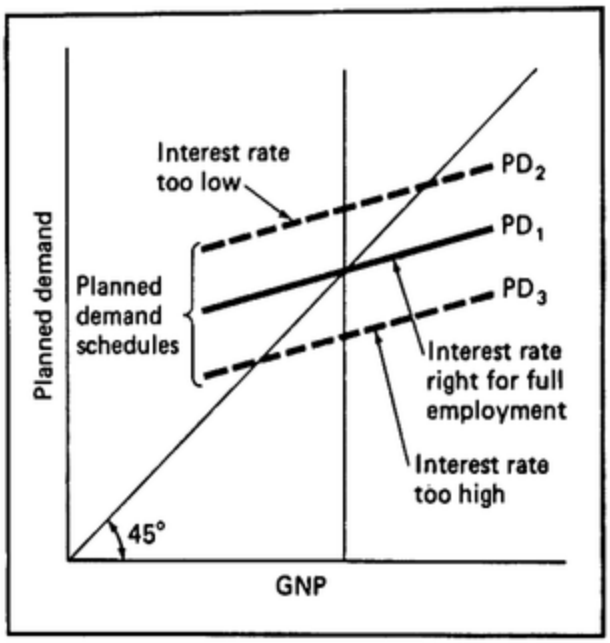


Figure 9 Planned demand and full employment

To sustain a level of GNP at which the unemployment rate equals the natural rate, the interest rate must settle at the right level, so that the planned demand schedule intersects the line of equality at Point *A*. Any lower interest rate will produce too low an unemployment rate. Any higher interest rate will produce too high an unemployment rate.

Or does it? Suppose the nominal money supply is too large and the interest rate too low. Planned demand will cut the line of equality to the right of full-employment GNP, as shown by PD_2 in Figure 9. Actual GNP will be higher than full-employment GNP, and the unemployment rate will, therefore, be lower than the natural rate. There will be *inflation*. If the Fed does not increase the monetary base, rising prices will destroy real money, and the real money supply schedule will shift to the left until the interest rate is high enough to end the inflation.

Similarly, suppose that the nominal money supply is too small and the interest rate is, therefore, too high. Planned de-

TEXT BOOK LIBRARY
Acc No 15358
Date

mand will then cut the line of equality to the left of full-employment GNP, as shown by PD_3 in Figure 9. GNP will be lower than its full-employment level, and the unemployment rate will, therefore, be higher than the natural rate. There will be *deflation.* (Economists who think the economy tends toward full employment think prices are flexible downward as well as upward.) Falling prices will increase the real money supply relative to GNP until the interest rate is low enough to end the deflation.

These arguments conclude that whatever the nominal money supply is, the price level will always adjust to create a real money supply consistent with full-employment GNP. The nominal money supply *in the long run* just determines the price level, and the money wage rate. Real GNP, the unemployment rate, and the interest rate are independent of the nominal money supply.

What, then, determines the equilibrium interest rate in the long run? In an economy without foreign trade or government, the long-run equilibrium rate equates planned investment and saving at full employment. As economists used to say, interest is determined by thrift and the productivity of capital. In the world as it is, the long-run equilibrium interest rate has to equate planned surpluses and deficits at full employment. It is determined by the forces that shape the trade balance, the government deficit, the business deficit, and household saving.

Be sure you understand that this is an argument about *the long run,* and a very "iffy" argument at that. Interest is determined by productivity and thrift only if the economy tends, on average, to be precisely balanced at full employment. That is a big "if." No one, not even the most ardent monetarist, doubts that monetary changes determine interest and GNP changes in the short run, over the business cycle. The claim that money is "neutral" in its effects on GNP, employment, and interest is at best a statement about what is true of averages over time, not what is true all the time.

Summary

This has been a long and difficult chapter. It is one of the most important in the book for someone who wants to understand how the economy fits together as a whole. The main points that you need to remember are the following:

1. Interest rates on various kinds of assets tend to move up and down together, as investors compete away differences that do not reflect risk, tax status, maturity, and similar considerations.
2. The equilibrium level of interest is determined by the supply of and demand for money. The supply curve shifts in response to changes in Federal Reserve policy. The demand curve shifts in response to changes in GNP.
3. Changes in the rate of interest lead to changes in the velocity of money in the same direction because interest is the cost of holding money. A change in the interest rate allows GNP to change relative to the money supply.
4. The Fed's control over the monetary base, reserve requirements, and the discount rate gives it the power to affect interest rates by shifting the money supply curve.
5. Changes in the interest rate lead to changes in planned demand in the opposite direction. The main direct effects are on residential and business investment.
6. The Fed's influence on the interest rate gives it an indirect power over GNP.

7. Since GNP affects the demand for money (through the GNP effect) and the interest rate affects planned demand for goods (through the interest effect), interest and GNP are mutually determined. This pattern of two-way causation is called the monetary feedback. It is a source of stability in the economy.
8. The Keynesian-monetarist debate centers largely on the strength of the monetary feedback. The monetarists think it is strong. It follows from this that autonomous changes in the financial market have a major effect on GNP, and autonomous changes in the planned demand for goods are largely canceled by the monetary feedback. The Keynesians think it is weak and therefore draw opposite conclusions about the relative importance of disturbances in the financial and goods markets.
9. Changes in the price level change the real money supply relative to GNP. Inflation reduces the real money supply and drives up the interest rate. Hence, inflation tends to bring itself to an end unless the Fed continually offsets the effect of rising prices by increasing the nominal money supply.
10. In times of persistent expected inflation, nominal interest rates may be relatively high. But real interest rates, adjusted for the expected rate of inflation, may be too low to curb the inflation. Again, the ultimate responsibility for prolonging the inflation belongs to the Fed, for allowing the money supply to increase too rapidly.
11. In an economy that tends toward full employment, money is neutral in the long run, affecting nominal variables, but not real variables.

Key concepts

Rate of return
Cost of funds
Rate of interest
Velocity of money
Interest effect
GNP effect
Monetary feedback
Monetary policy
Keynesianism
Monetarism
Crowding out
Real stock of money
Real rate of interest
Neutrality of money

Questions for review

1. Considering the characteristics of the following assets, choose the asset in each pair that you would expect to have the higher yield. Explain your choice.
   a. a savings account or a money market fund
   b. a municipal bond or a corporate bond of equal risk
   c. a checking account or a savings account
2. Explain how the concept of *opportunity cost* is relevant to people's decisions about how much money to hold.
3. Suppose that your income increases. Is it possible for you to increase your level of transactions without increasing your average holding of money balances? Explain.
4. a. What determines the *height* and *slope* of the demand schedule for money?
   b. What determines the *height* and *slope* of the supply schedule of money?

c. How will the following changes affect the position or slope of the money supply and demand schedules?
   i. The planned demand schedule shifts up.
   ii. The Fed sells securities.
   iii. Velocity becomes more sensitive to changes in the rate of interest.
   iv. Military expenditures decrease, causing equilibrium GNP to fall.
   v. The reserve requirement is lowered.

5. Why is it harder to identify the relation between interest rates and business investment than between interest rates and residential investment?

6. a. Explain what is meant by the monetary feedback. In doing so, point out the difference between the interest effect and the GNP effect.
   b. Does the operation of the monetary feedback make the economy more stable or less stable? Explain.
   c. Each of the following situations will trigger the monetary feedback. Determine whether the monetary feedback in each case will begin with the interest effect or the GNP effect.
      i. The Fed sells government securities.
      ii. Consumers begin to save more at every level of income.
      iii. The Fed reduces the reserve requirement.
      iv. Congress legislates a decrease in tax rates.
      v. Reserve requirements are increased.

7. Identify each of the following views as a *monetarist* or *Keynesian* view:
   a. Velocity is almost independent of the interest rate.
   b. Changes in the monetary base will have very little impact on GNP.
   c. Changes in the interest rate have a very large impact on planned demand.
   d. A shift in the demand for money schedule will have a relatively large impact on interest rates.
   e. The GNP effect is fairly weak.
   f. Instability in financial markets causes great instability in the goods market.
   g. The monetary feedback will not stabilize GNP in cases of autonomous changes in planned demand.

8. a. Describe the difference between the nominal stock of money and the real stock of money. Which measure will give a clearer indication of changes in velocity?
   b. What is the difference between the nominal rate of interest and the real rate of interest?

9. The Fed is actively promoting inflation if it allows the monetary base to increase, keeping pace with inflation. True or false? Explain.

30

The Institutions, Goals, and Strategies of Stabilization Policy

As you read and study this chapter, you will learn:

- which parts of the federal government are responsible for trying to stabilize output, employment, and prices
- how the government sets its stabilization goals
- what the costs are of a failure to meet these goals
- some of the overall strategies that the government might pursue as it tries to meet its goals

Occasionally a commercial product or work of art is so successful that its name becomes a common noun. Kleenex, Scotch Tape, and Catch 22 are good examples. You can probably think of others. Not so long ago there was a cartoonist in this country named Rube Goldberg. His pictures were pure fun, without any political message. They showed very complicated devices for doing very simple things. For decades, any ridiculously complex contraption was known as a "Rube Goldberg device" (see Figure 1). Everyone knew what this meant, even people who hadn't the faintest idea who Rube Goldberg might be.

If you have ever studied the structure of American government, especially the "separation of powers," you may have been reminded of a Rube Goldberg machine. The institutional messiness is especially obvious when it comes to ***stabilization policy***—coordinated government action designed to smooth out the business cycle and to limit changes in the overall price level. The conduct of stabilization policy is the subject of this chapter and

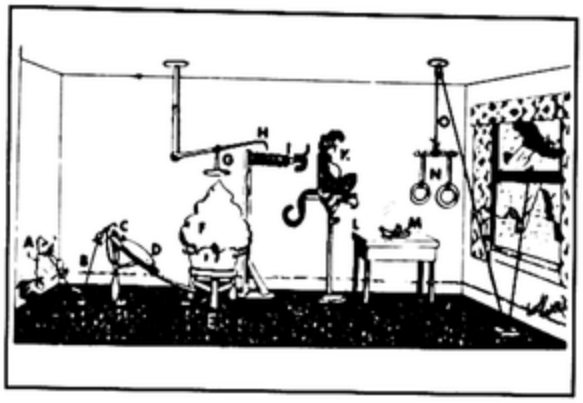


Figure 1 A Rube Goldberg device

The professor takes a pill and dopes out a device for closing the window if it starts to rain while you're away. Pet bullfrog (A), homesick for water, hears rainstorm and jumps for joy, pulling string (B), which opens catch (C) and releases hot water bag (D), allowing it to slide under chair (E). Heat raises yeast (F), lifting disk (G), which causes hook (H) to release spring (I). Toy automobile-bumper (J) socks monkey (K) in the neck, putting him down for the count on table (L). He staggers to his feet and slips on banana peel (M). He instinctively reaches for flying rings (N) to avoid further disaster and his weight pulls rope (O), closing window (P), stopping the rain from leaking through on the family downstairs and thinning their soup.

Source: Copyright © RUBE GOLDBERG, distributed by King Features Syndicate, Inc. By permission.

of the two that follow it. Nearly every federal agency gets into the act, whether it wants to or not. Government spending and taxing, which involve Congress and various branches of the Executive, are particularly chaotic. Control over the money supply, by contrast, is vested in a single agency, the Federal Reserve. Even though changes in the money supply should obviously be coordinated with the federal budget, the Federal Reserve is legally independent of the Executive and Congress in its day-to-day operations. This fragmented pattern of responsibility is about as sensible-looking as a camel, a "horse designed by a committee." We could not begin to describe it fully in a few pages. But to give you a concrete image when you think about policymaking, we will first briefly discuss the major institutions and their responsibilities. No specific attention will be given to the Federal Reserve, since you have just studied it extensively. The rundown is restricted to the institutions that plan *budgetary or fiscal policy.*

The conduct of fiscal policy

It was widely expected during the latter years of the Second World War that there would be a postwar slump, perhaps a resumption of the Great Depression. Various government officials and economists argued that the risks of reconversion to peacetime production would be much smaller if the federal government took formal responsibility for maintaining full employment. Their conception gained limited support in the administration and Congress, and in January 1945 (before the end of the war), a group of senators introduced a bill known as the "Full Employment Act of 1945," which declared it to be the will of Congress that "all Americans able to work and seeking work have the right to useful, remunerative, regular, and full-time employment." It is hard to imagine a more direct statement. The bill that was finally passed by Congress and signed by President Truman in 1946 was far weaker than the original bill, which had been strenuously opposed as "socialism" by the business community and others. For better or worse, the Employment Act of 1946 defined the outlines of postwar policy planning.

The Employment Act of 1946 and the Council of Economic Advisers

The declaration of policy that made up Section Two of the act reads as follows:

> The Congress hereby declares that it is the continuing policy and responsibility of the Federal Government to use all practicable means consistent with its needs and obligations and other essential considerations of national policy with the assistance and cooperation of industry, ag-

riculture, labor, and state and local government, to coordinate and utilize all its plans, functions, and resources for the purpose of creating and maintaining, in a manner calculated to foster and promote free competitive enterprise and the general welfare, conditions under which there will be afforded useful employment, for those able, willing, and seeking to work, and to promote maximum employment, production, and purchasing power.

If you wrote this sentence, you could expect to be sent back to Freshman Composition. It looks so silly because it represented a compromise among powerful interests whose views could not be reconciled in any consistent way. These inconsistent views were just listed in alternating phrases, a "one for me and one for you" approach.

Despite its Alice in Wonderland declaration of policy, the act did at least one important thing: it created a three-member ***Council of Economic Advisers*** (CEA), to be appointed by the President with the consent of the Senate, "each of whom shall be a person who, as a result of his training, experience, and attainments, is exceptionally qualified to analyze and interpret economic developments . . . in the light of the policy declared in Section Two." This group, which in practice has been made up of professional economists, was "to assist and advise the President . . . to gather timely and authoritative information . . . to appraise the various programs and activities of the Federal Government . . . to develop and recommend to the President national economic policies . . . to avoid economic fluctuations or diminish the effects thereof, and to maintain employment, production, and purchasing power." In other words, the act created an agency that had the responsibility for formulating stabilization policy and advising the President on how to maintain prosperity. Over the years, this institution has gained influence. Most presidents have sought out able advisers and have relied on them heavily, particularly in recent decades.

You must not imagine, however, that the Council *makes* stabilization policy. It advises the President, testifies before congressional committees, and consults with other Executive agencies and with the Federal Reserve. But it has neither ad ministrative functions nor the staff to do more than give advice about legislation.

The Office of Management and Budget and the Cabinet-level departments

Overall responsibility for the budget that the President proposes to the Congress each year resides in the Office of Management and Budget (OMB). Like the Council of Economic Advisers, the OMB is part of the Executive Office of the President. The OMB and the CEA occupy the same building and consult about the stabilization aspects of taxes and expenditures.

The OMB also works closely with the Cabinet departments (Labor, Agriculture, Commerce, Defense, Education, Housing and Urban Development, Interior, etc.). These agencies are responsible for administering federal programs and often initiate new ones. The content of government spending is largely determined by the size and nature of these programs. In estimating tax revenues, the OMB must rely heavily on projections coming from the Treasury Department, which is also responsible for formulating proposed changes in tax laws. Thus, the overall process that determines budget policy is a complex interaction among many agencies.

The Congress

The administration's budget is just a very elaborate recommendation to Congress. As you can well imagine, Congress invariably has its own ideas about what needs to be done, and legislating the budget is a lengthy and complicated process. Expen-

ditures are shaped in the House and Senate appropriations committees, and new programs must be sent to other committees for hearings and analysis. All new tax legislation is initially considered in the House Ways and Means and Senate Finance committees. All bills must be passed by the full House and Senate. If the two houses pass different bills, the differences must be ironed out in conference committees. Finally, the President must sign the bill into law. The whole process had become so drawn out by the 1970s that most yearly appropriations bills had not been passed in time for the July start of the fiscal year. This meant that the agencies were usually operating under "continuing resolution," a congressional authorization to continue to spend at the rate appropriated in the previous fiscal year. To catch up, Congress finally redefined the fiscal year to begin on October 1. To do this, it simply declared July–September 1976 a three-month gap in fiscal time. Many of us would find it a relief to be able to do this now and then.

Because of the complexity of the budgeting process, it is not easy to achieve precise timing in fiscal policy. However, the system works more coherently in practice than it looks on paper, partly because most of the time the departments, the CEA and the OMB, the President, and the Congress share broadly similar values and therefore a common view of what needs to be done. But don't expect anything to happen very fast unless there is an obvious emergency.

The goals of stabilization policy and the costs of instability

The 1962 Economic Report

The goals of stabilization policy expressed in the Employment Act of 1946 are to "promote maximum employment, production, and purchasing power in such a way as to foster and promote free competitive enterprise and the general welfare." This is a tall order. It is also not very specific. Like the civil and criminal law, it has acquired its content as it has been applied.

The task of giving substance to the Act has largely been taken on by the Council of Economic Advisers. Like any new institution, the Council had to define its own role in government as it developed. The Council first attained major influence in policy formulation during the Kennedy and Johnson administrations. A succession of active CEA chairmen (see Figure 2)—Walter Heller (1961–1964), Gardner Ackley (1964–1968), and Arthur Okun (1968–1969)—were largely responsible for impressing upon the government the necessity of coordinating various policies, so that their overall impact would be consistent with prosperity.

Policy guidelines

Suppose that you were appointed to advise on stabilization policy. What sorts of things would you think about before your first meeting with the President? How would you describe the stabilization problem that he or she would have to face?

One possible agenda would be to propose an overall goal, then to pinpoint a set of specific quantitative targets, and finally to formulate a policy program for hitting those targets. If you shared the values of presidential advisers during the 1946–1980 period, your overall goal would be to achieve the lowest average rate of unemployment consistent with long-run price stability. Your specific quantitative targets for unemployment and GNP would depend on your particular understanding of how the economy works. Thus, you might even set different goals from those set by other advisers who shared your values. Finally, your specific policy program would depend not only on your targets but also on

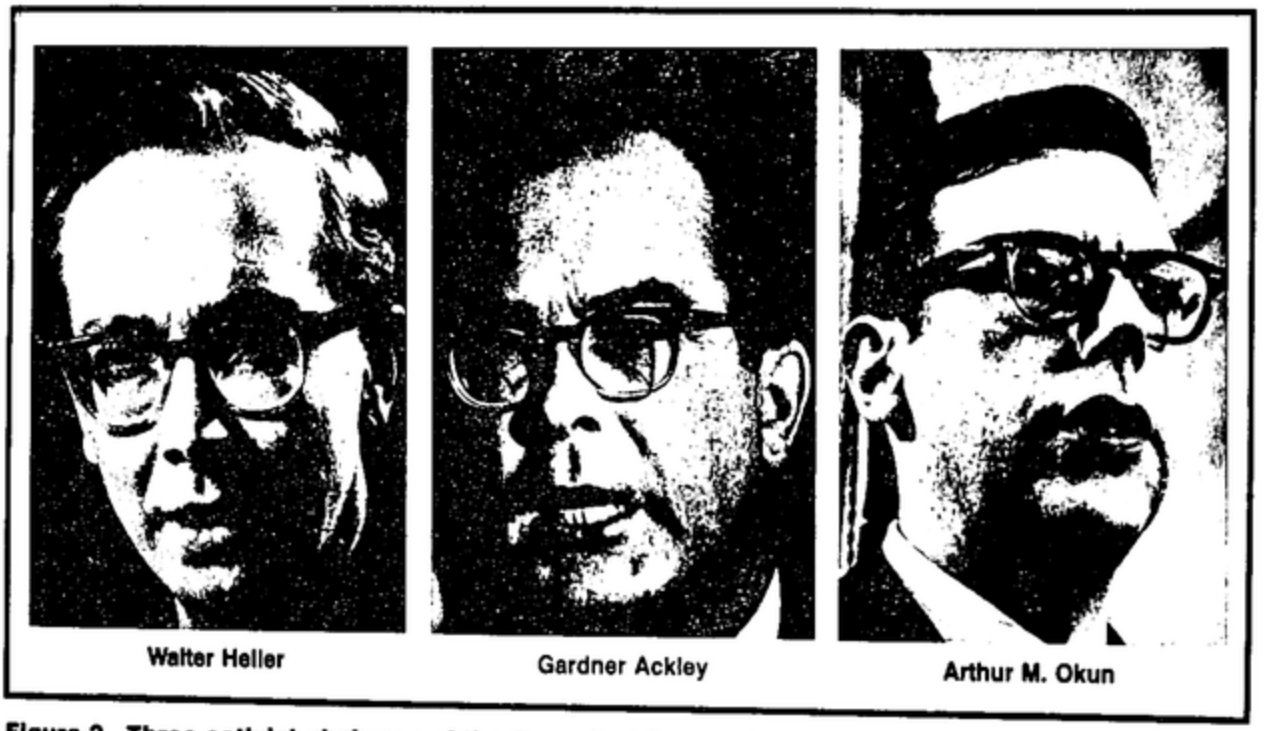


Figure 2 Three activist chairmen of the Council of Economic Advisers

your assessment of the current situation and of the effectiveness of various alternative policies.

Potential GNP and the gap

The CEA that took office along with John Kennedy in 1961 shared with its predecessors a commitment to the goals of high employment and stable prices. But it was the first Council to set a specific quantitative target for stabilization policy to aim at. It called this target ***potential GNP.*** This concept first appeared in the Council's *Report* for 1962, which largely focused on unemployment. By looking at the post–World War II record, the Council concluded that the economy could sustain a 4 percent unemployment rate with only a moderate rate of price increase. The unemployment rate had last been 4 percent in mid-1955. Between then and 1962, the combined rates of productivity and labor force growth had been about 3½ percent a year. So the Council drew a 3½ percent trend line through the GNP of mid-1955, arguing that the actual GNP could have followed this path without serious inflationary consequences. It called this trend *the path of potential GNP.* It is essentially the low-inflation boundary you encountered in an earlier chapter. A graph of potential and actual GNP for 1952–1980 is shown in the top part of Figure 3.

The shaded areas in Figure 3 flag periods when actual GNP was above or below potential. The word "potential" was never intended to imply a rigid ceiling that the economy could not pierce, but rather a "redline" level that could not be exceeded for long without serious inflation. You can find instances in Figure 3 when GNP was above its potential, particularly from 1965 through 1969, during the Vietnam War. Predominantly, however, GNP has fallen short of potential over the past three decades. When it has, there has been substantial unemployment.

To underscore the relationship between unemployment and GNP, the 1962

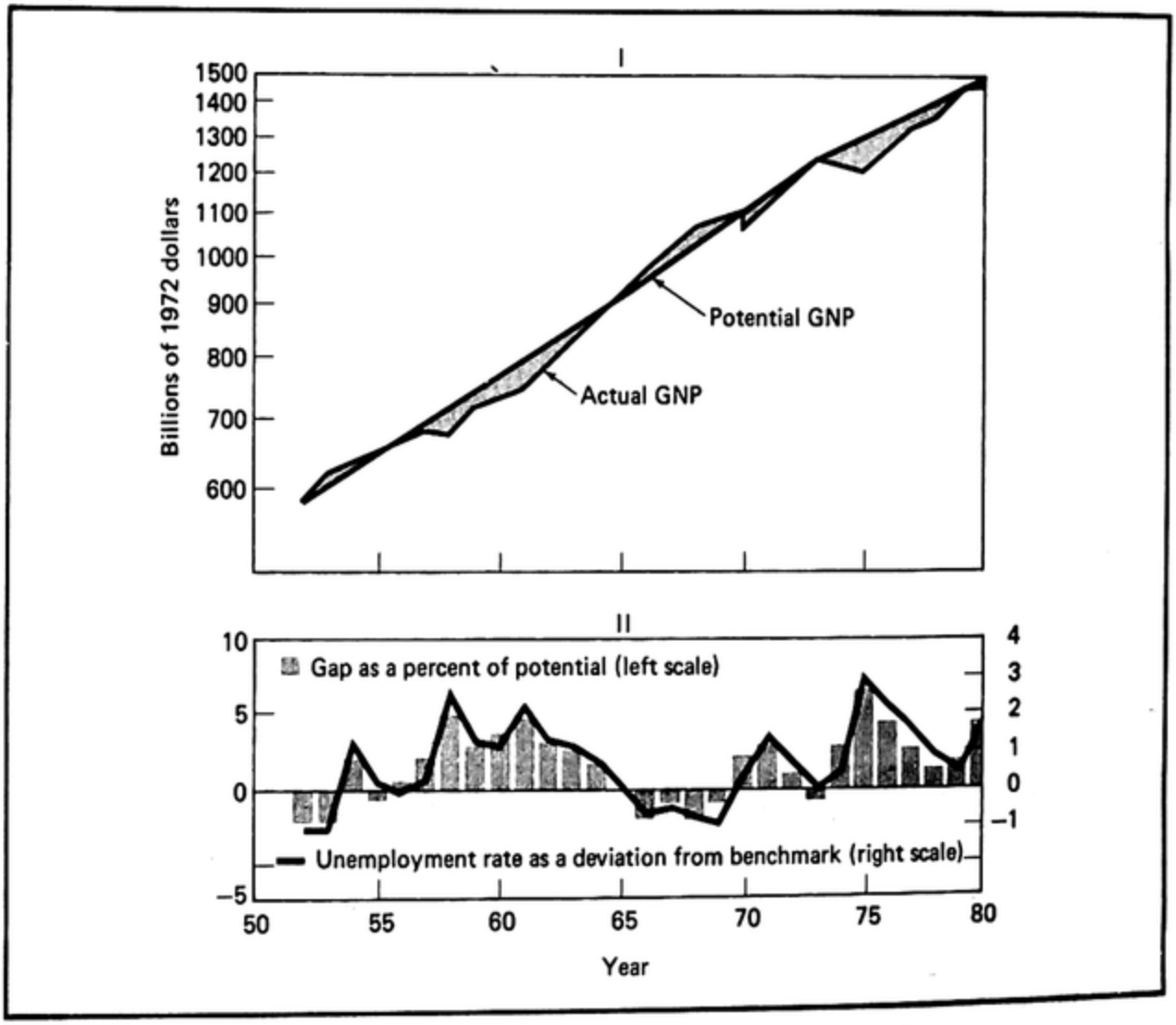


Figure 3 Potential and actual GNP, unemployment, and the gap 1952–1980

The line labeled "Potential GNP" marks the Council's estimates of the maximum levels of GNP that are consistent with price stability. As Panel I shows, actual GNP has occasionally been above potential, but is usually below it. Panel II shows that when actual GNP is below or above potential, the unemployment rate is above or below the benchmark rate that corresponds to potential GNP.

Source: *Economic Report of the President.*

Report defined a ***gap***—the difference between potential and actual GNP. When potential is above actual, the gap is positive. When actual is above potential, the gap is negative. The *Report* then compared the gap to the unemployment rate. You can see this comparison in Panel II of Figure 3. The gap is measured as a percentage of potential GNP. The unemployment rate is measured relative to the low-unemployment *benchmark*. Until about 1968, this benchmark was the 4 percent unemployment rate used in the 1962 *Report*. However, Council *Reports* during the 1970s argued that because of a slow drop in the percentage of the labor force made up of "mainstream" workers, among whom unemployment is rare, the benchmark had risen since the 1960s, reaching about 5 percent by 1979. Both liberal and conservative Councils have agreed on the significance of this trend.

The correlation between the gap and the unemployment rate is impressive.

Judging from the figures, a gap equal to 2½ percent of potential GNP raises the unemployment rate 1 percentage point above its benchmark level. Put differently, every 1 percentage point drop in the unemployment rate means an increase in GNP equal to 2½ percent of potential.

If you think of potential GNP the way the 1962 Council did, you may picture it as a tightrope. The goal of stabilization policy is to keep the economy on the tightrope. If it falls off—no matter on which side—there are costs to be paid. A negative gap, with GNP above potential, leads to demand-pull inflation. A positive gap, with GNP below potential, leads to unnecessarily high unemployment. The rewards for successful stabilization policy are measured by the costs of failure.

The gap as lost output

Whenever the gap is positive, it tells us how much GNP we waste by failing to produce at a level that could be sustained without serious inflation. To put the amounts in perspective, suppose that instead of falling 7.7 percent short of our potential in 1975, we had instead produced that additional output and distributed it to the underdeveloped countries of Africa. They could have doubled their living standards. Because we failed to use this potential in any way at all, a staggering volume of output was simply lost forever. On average, we wasted about 3 percent of our potential output in this way during the 1970s. Imagine what these goods would have meant to the truly poor people of the world.

The gap as wasted resources

Wasted output is just one face of a coin whose other face is unemployment. During the 1970s, while we were wasting about 3 percent of our potential output each year, the unemployment rate averaged about 1.3 points above the Council's benchmark. Most of this unemployment represented useful work that people would have performed if jobs had been available. Moreover, the wasted labor power that shows up in the unemployment statistics seriously understates the total loss. Table 1 shows how the 1961 Council allocated the gap among its various components. Only 40 percent of the waste of labor power shows up as measured unemployment. The rest is ***hidden unemployment*** in its various forms. One portion (10 percent) fails to appear in the unemployment statistics because people who would like jobs drop out of the labor force when they cannot find them. As long as they are not actively seeking work, their unemployment is not counted. A second 10 percent is attributed to involuntary shortening of the work week of those still on the job. The remaining 40 percent is assigned to reduce productivity of those actually on the job. Much of the work force is engaged in overhead activities: administration, sales, clerical work, and supervision. When firms are working at less than capacity, this overhead is spread over a relatively small output, and labor is less productive than it would be at full utilization.

These guesses at hidden unemployment are rough, but educated. Everyone who has looked at the problem has concluded that measured unemployment greatly underestimates the waste of labor power during recessions. To relate it to

Table 1 Allocation of the 1961 GAP: Losses attributed to measured and hidden unemployment

Share of the gap attributed to:	Percent
Measured unemployment	40
Hidden unemployment	
Reduced labor force	10
Reduced hours of those working	10
Reduced productivity of those working	40
	100

Source: *Economic Report of the President.*

your own experience, think about retail stores after the Christmas rush is over, with their idle clerks and anxious managers. Then picture the same scene all over the country, in industry and services as well as trade. Picture also a shortened work week and people going back to school or staying home, who would be looking for work if they thought they could find it. Then you will understand what is meant by hidden unemployment.

The cost of exceeding potential output

Greek mythology has given us one of our most overused metaphors: to be caught "between Scylla and Charybdis," a rock and a whirlpool that threatened ships in the straits that separate Sicily from mainland Italy. To be caught between Scylla and Charybdis is to face grave dangers on both sides—going too far to avoid one of them exposes you to the other.

American folk speech expresses the same hard choice as being "between a rock and a hard place." Anyone discussing both unemployment and inflation finds it hard to resist using one of these metaphors. The path of potential GNP seems to lead safely between the rock of underemployment and the hard place of overemployment. The costs of hitting the rock are clear enough. They are unemployed people and lost output. But what about the hard place above potential GNP? What are the costs of going in this direction?

The word "potential" stands in the way of understanding these costs, since to exceed your potential generally means to do something almost heroic, to do *better* than anyone might have predicted. This is not how you should interpret exceeding potential GNP. It is more like "getting into deep water, in over your head." It is neither difficult nor meritorious, just dangerous, and probably stupid.

The danger, of course, is inflation. We have already noted that the Council's path of potential GNP is essentially the low-inflation boundary we used earlier. Exceeding potential GNP, therefore, means crossing over into the territory of the inflationary process. To do this briefly means gaining extra output and employment at the cost of temporary inflation. This happened in the mid-1950s, and caused alarm, but no lasting damage. Exceeding potential for a long period, as we did from 1966 through 1969, is entirely different. As you know, this contributed to the persistent inflation that continued through the 1970s and is probably still continuing as you read this book. ***Putting an end to this kind of inflation imposes real and unrecoverable costs.***

The only sure way to break the momentum of a persistent inflation is to break the expectations that keep it going. Remember that the hallmark of persistent inflation is a general belief that it will continue. When suppliers and demanders both expect inflation, they act in ways that make inflation continue. As long as their expectations prove correct, they continue to act on them. Most economists think that because these inflationary expectations are learned by experience, they can only be unlearned by experience. This means that prices and wages must somehow be made to rise less rapidly than people expect. The most obvious way to do this is to go through a recession or depression that is severe enough to stabilize prices even though people expect them to rise. If this analysis is correct, then to put an end to persistent inflation, we must deliberately squander productive potential and resources. These are the real and unrecoverable costs of inflation.

Stopping persistent inflation: The costs

To help yourself focus on these costs, think back to the Phillips curve. As you may recall, the relationship between inflation and unemployment is not fixed. The Phillips

curve shifts up and down according to people's expectations about future rates of inflation.

If 6 percent is the natural unemployment rate, either of the two circled points in Figure 4 is sustainable, in the sense that expectations are borne out by what happens. One has 6 percent unemployment and 8 percent price inflation, the other 6 percent unemployment and price stability. Which would you choose? Most people would choose the bottom one, since it seems to offer greatly improved price performance at no unemployment cost.

In fact, this "something for nothing" is an illusion, at least for an economy that is on the top curve. Although it is costless *to be* on the bottom curve rather than the top one, it is not costless *to get there*. Satellites can stay in orbit without burning their engines, but it can be extraordinarily expensive to get them from one orbit to another.

The cost of getting from the top curve to the bottom one is measured in lost output and wasted resources. Quite simply, it is the cost of going through a recession long and serious enough to break the expectations that stand in the way of price stability. No one really knows how long this process would take. We do know that each extra percentage point of unemployment lowers the *actual* rate of inflation by about ¾ of a percentage point, or at any rate it did in the 1950s and 1960s. But we don't know how fast expectations adjust to the actual rate because we have very little data on inflationary expectations.

Some economists believe that the general public forms ***rational expectations***. On the average, it *correctly* assesses the implications of current events and policies. Hence, the public can tell a policy that will stop inflation from a policy that will not. Providing that the government is clearly

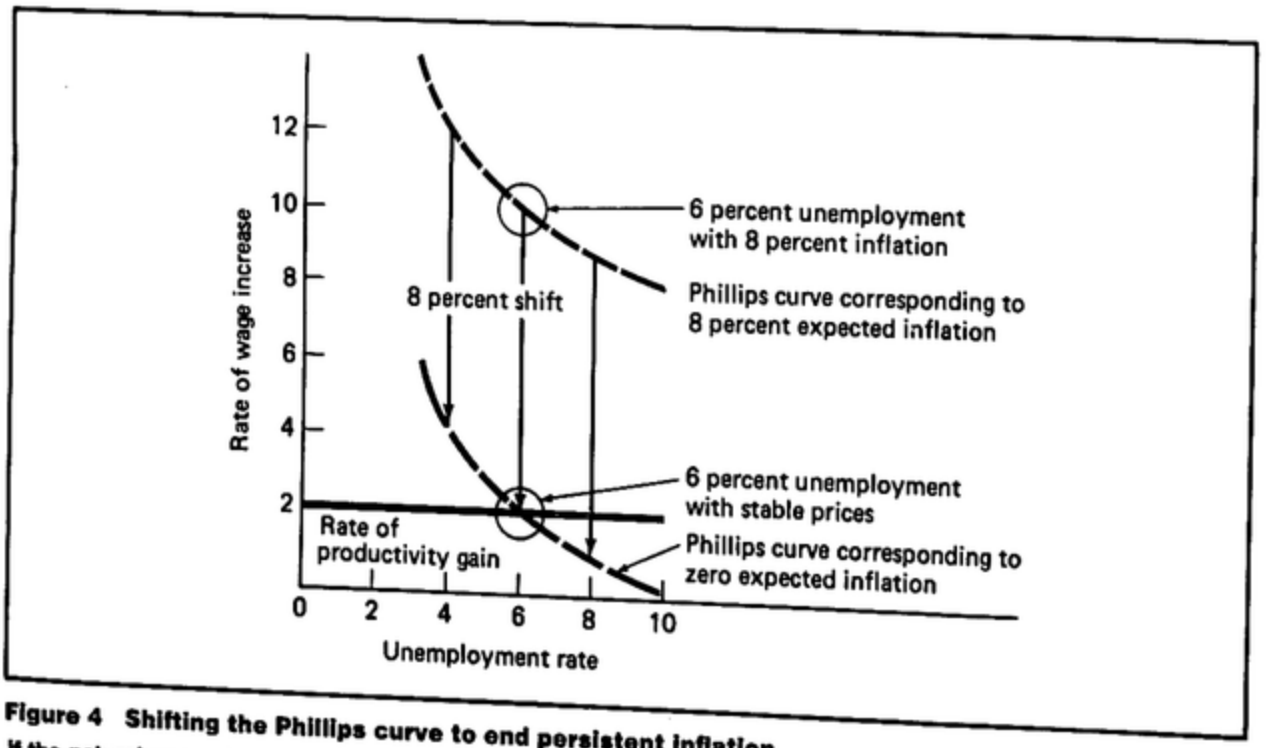


Figure 4 Shifting the Phillips curve to end persistent inflation

If the natural unemployment rate is 6 percent, either of the two circled points in the diagram is sustainable. The bottom one is more desirable, since it has a lower inflation rate and the same unemployment rate. But if the economy is situated at the top point, it must go through a period of high unemployment to get to the bottom point. Moving down, therefore, has real costs.

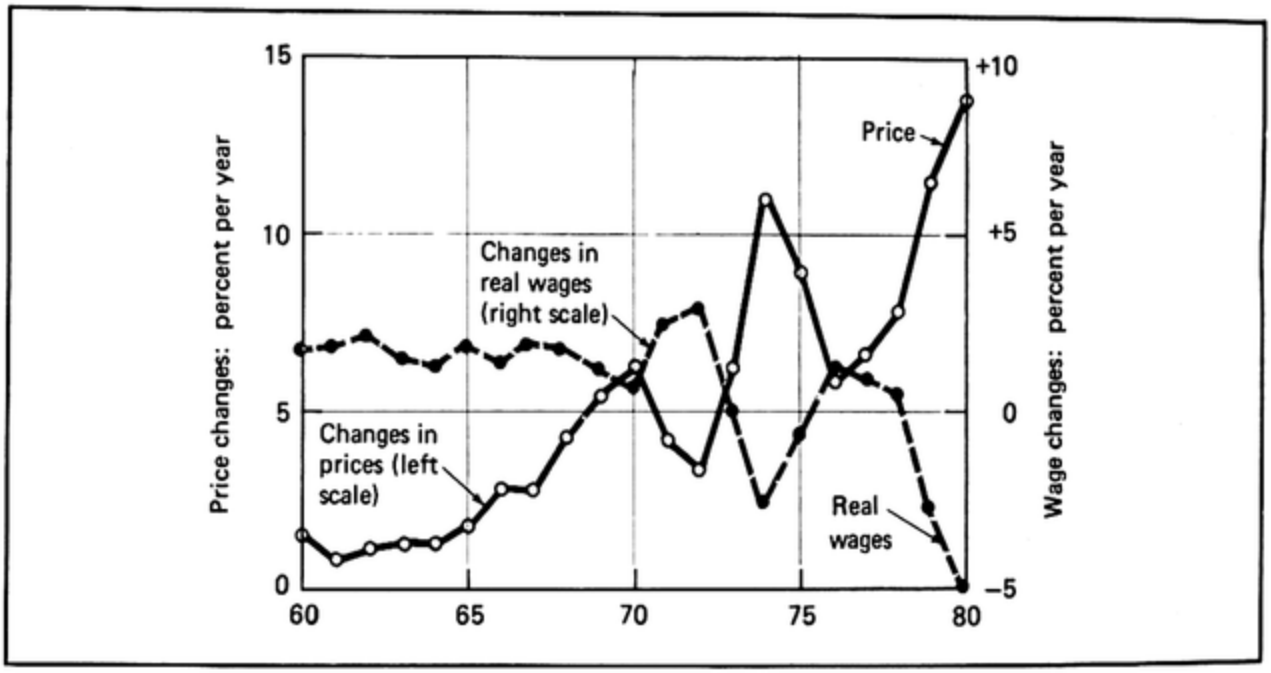


Figure 5 The consumer price index and real hourly earnings 1960–1980

Every peak in the inflation rate shows up as a trough in the rate of change in real hourly earnings of production workers.

Source: *Economic Report of the President.*

determined to pursue an anti-inflation policy to the bitter end, inflationary expectations will "melt away like the morning mist," as Reagan's budget director said. These economists believe that persistent inflation can be speedily halted by putting tough-minded people like themselves in power.

This group is optimistic, in its tough-minded way. The pessimists, however, think that expectations have to be learned through experience. They believe that it may take years to break inflationary expectations, years during which the economy is so depressed that prices rise slowly even though people expect them to rise rapidly.

Living with persistent inflation: The costs

As the inflation that began in the late 1960s gathered momentum, there was an outcry from wage earners, who saw their money-wage gains largely eaten up by price increases. The "double-digit" (more than 10 percent) inflations of 1974 and 1979–1980 amplified this cry until it drowned out much of the rest of the news. If you look at Figure 5, you can see what the outcry was about. Every peak in inflation coincides with a trough in the rate of growth of real hourly earnings, measured in dollars of constant purchasing power. In years of double-digit inflation, real wages fell sharply.

Does this mean that the inflation shifted real income away from wage earners and toward recipients of other forms of income? It did not. The share of labor income in GNY was larger (61.3 percent) in the highly inflationary 1970s than it was in the less inflationary 1960s (when it was 58.6 percent). How can this be? If rapid inflation lowers real wages, shouldn't persistently rapid inflation distribute income away from labor?

No, because persistent inflation comes to be anticipated: Only the peak rates of inflation come as a surprise. They lower real wages sharply because they are unan-

ticipated and not built into the wage determination processes. Anticipated inflation, on the other hand, is taken into account in wage negotiations in both organized and unorganized labor markets. Therefore, inflation does not permanently reduce labor's share of national income, although it may hurt small, vulnerable groups of wage earners. Real wages grew very slowly during the 1970s, only 0.2 percent a year compared with 1.9 percent a year during the 1960s. But since productivity gains were also much lower in the 1970s, there was no massive shift toward profits.

In thinking about the costs of inflation, it is important to distinguish between *anticipated* and *unanticipated inflation*, as the above example makes clear. Unanticipated inflation substantially redistributes wealth and income. In the very short run, wage earners, pensioners, and lenders lose. Profit recipients, debtors, and owners of real property gain. Over the longer haul, however, money wages, interest rates, and pensions come to incorporate allowances for the rate of inflation, and redistribution is much more limited. If cost-of-living adjustments were built into all contracts, as they are in many wage bargains negotiated in the unionized industries, they would eliminate much of the redistributive effect of inflation.

Under a fully anticipated inflation, the structure of relative incomes and prices would look much the way it would under conditions of price stability: *All prices and incomes* would have a built-in inflation factor, common to all. Nominal interest rates would be high, but real rates would be normal. Only money holding would be systematically penalized, and the payment of inflation-linked interest on checking accounts would offset much of this penalty.

This way of protecting people against the redistributive costs of inflation is known as ***Indexing*** because it ties all contractual payments to some price index such as the CPI. If indexing were sufficiently general, the installments you pay on your tuition would consist of a contractual amount, agreed to when you enroll, plus an amount that varied with the rate of inflation since you enrolled. Your teachers' salaries and the price of football tickets would be similarly determined. No matter what the rate of inflation, the ratios among tuition, salaries, and ticket prices would remain where they were set at the beginning of the school year. The interest and principal on your student loan would also be indexed, so that a 5 percent rate of interest when prices are stable would become 15 percent at an inflation rate of 10 percent. In effect, this would fix the *real rate of interest*, so that it would not vary inversely with the rate of inflation.

This must sound a lot like price control to you, but it is not. Prices that are determined daily on competitive markets would not need to be indexed. Neither would those that can be changed at will by sellers. Only *contractual* payments would need to be indexed, and the base level of the contract (the first tuition installment, for example) would be agreed to by the contracting parties in the usual way. Only changes over the life of the contract would be determined by the index. Many people like this idea, and the inflation of the 1970s increased the popularity of indexing. Great uncertainty about future inflation rates makes people uncomfortable about entering into long-term contracts. Indexing reduces uncertainty and encourages people to behave as if prices were generally stable.

Indexing has one great drawback, however. In a fully indexed economy, inflation would spread like wildfire. Every increase in the price level would be followed automatically and almost immediately by additional price changes. As soon as the price increases showed up in the indexes, there would be still further increases. On the other hand, however, there would also

be a tendency for inflation to subside quickly. One thing that made the inflation of the 1970s persist even with high unemployment was the tendency for contract renewals to "catch up" to unanticipated inflation that had occurred during the life of the previous contract. To the extent that this pressure is relieved by indexing, current rates of price increase would more faithfully reflect current market conditions. Thus, although inflation would spread more rapidly under general indexing, it would also subside more rapidly.

So far, our discussion of the costs of living with inflation has centered on its redistributive effects. Unanticipated increases in the rate of inflation redistribute income and wealth away from some people, but they reward others. Of course, an unanticipated reduction in the rate of inflation would redistribute income in the other direction. Thus, to bring inflation under control would impose redistributive costs of its own. But redistribution is a *zero sum game*, in which every loser has a counterpart. By contrast, bringing inflation under control by causing a recession means lost output and wasted resources that are a cost to everyone. Does allowing inflation to continue cause similar *deadweight losses*, not offset by corresponding gains?

A substantial and conspicuous element in the inflation of the 1970s was the rise in the price of imported oil. This resulted in an international redistribution of income. From the point of view of the importing countries, this was indeed a deadweight loss. But this loss was a cause of the inflation, not a consequence. What we see is an unfortunate result of continuing inflation that could be avoided if prices were brought under control.

A fairly common argument is that inflation discourages saving, investment, and productivity growth. If prices are rising fast enough relative to interest rates, the return on physical assets is higher than the yield on financial saving. This accounts for the scramble in the 1970s to invest in precious metals, antiques, coins, works of art, and old comic books. The low return on savings also encourages people to buy consumer durables and housing rather than to make financial investments. This behavior limits the flow of funds to business through financial markets, and makes it more costly for firms to finance new productive plant and equipment. The result is lower growth.

If things worked out this way, it would be a real cost of inflation. In fact, they did not, at least in the 1970s. First, the consumer saving rate during the 1970s was normal, even somewhat larger than it was in the 1960s. Second, the flow of funds through the financial markets was enormous. Third, the real cost of funds to business was low rather than high for most of the decade. The same high rates of inflation that lowered the real return on lending also lowered the real cost of borrowing. Fourth, with all the unemployment and excess capacity, there was no shortage of resources for producing investment goods. Investment may have been depressed by the state of business confidence, but it was not crowded out by high real interest rates or insufficient capacity to produce investment goods. Thus, there were no obvious deadweight losses.

The best case against allowing inflation to persist is that it creates uncertainty and fear of worse inflation to come. In turn, this promotes industrial and political instability, brings on waves of strikes and shutdowns, and furthers the fortunes of political charlatans who promise to "restore order." There is a definite connection between the post–World War I European inflations and the rise of fascism. Even if this were not true, people detest inflation. To

the extent that it is possible, therefore, democratic governments should try to keep the price level under control.

Price and wage controls

After reading about the costs of living with inflation and of ending it, you may wonder why we don't make more extensive use of temporary *price and wage controls.* Couldn't they be put into effect just long enough to break the expectations of continued inflation? Then they could be removed, and the inflation ended without a prolonged or severe period of unemployment.

This is an appealing argument that has been advanced by some prominent people. ***Direct controls*** over wages and prices have curbed inflation in the past, particularly in wartime. They will probably be used again in the future. With the proper enforcement mechanisms, they will, in fact, stop wages and prices from spiraling.

Nonetheless, many economists oppose price and wage controls for several reasons. The most common is that it is virtually impossible to get the *structure* of controls right for long. Products or skills that are priced too low will be short in supply. Those that are priced too high will be supplied in quantities that exceed demand. Even if relative prices are just right at one time, they will soon get out of line, as costs and demands shift but prices do not. Shortages and surpluses will develop and grow as the allocation of resources gets out of line with demand. Eventually, the system will collapse because it is hopelessly inefficient and impossible to administer in any sensible way.

This argument is probably overstated. Most experience with controls comes from wartime, when shortages (usually coupled with rationing) result from *deliberate* decisions to produce some goods in small amounts, yet price them low enough so that most people can afford them if they can be found. It is hard to think of anything that would have to be deliberately priced in this way in a system of controls designed to halt persistent inflation.

A second argument, the claim that no system of controls could be ultimately effective, merits more careful consideration, however. People who argue this way point out that even when controls have worked for a while, they have been followed by mounting inflation as soon as they were lifted. Controls may postpone inflation, but they cannot halt it permanently. Note that this argument will be correct if enough people believe that it is correct. People, unions, and firms will try to protect themselves from inflation, and in doing so, will bring on just what they fear.

In fact, direct controls work rather well in socialist countries. American news media tend to sensationalize the failures of socialist economic systems, but along with their obvious failures, they do succeed most of the time in controlling consumer prices. The reason for this is that controls are accepted as a normal and permanent state of affairs. Socialist bureaucrats are accustomed to administering them and have faith in their rationality. Similar controls are less effective in our economy because everyone views them as unfortunate and temporary expedients, like braces on your teeth.

Given these doubts about controls in the United States, only a deliberately engineered recession seems likely to halt persistent inflation. Whether this is a desirable policy depends on whether the benefits outweigh the costs. When the benefits go to people who do not have to bear the costs, it is not even clear whose scale should be used to weigh one against the other. A tenured professor whose salary consistently lags behind inflation has a lot

to gain from stable prices. An auto worker with no seniority has a lot to lose from a deliberately engineered recession.

There is no point in imagining that a decision to stop inflation or let it continue is anything but political and difficult. It may help you to make up your own mind if you are better informed about the costs of continuing inflation versus the costs of controlling it, but the facts will not make a decision for you. Indeed, the facts themselves are not all that well known.

Strategies of stabilization

This chapter has made two general points thus far:

1. In the American political-economic system, the responsibility for regulating the money supply is lodged in an independent Federal Reserve, and the responsibility for budgeting is shared among many executive agencies whose collective decisions are greatly modified by Congress.
2. The problem of stabilization is one of guiding a largely private economy along a narrow path between the territory of inflation and the wasteland of underemployment and unused potential.

You might suppose that everyone who had studied the stabilization problem would long ago have concluded that what we need most of all is greater institutional flexibility and coordination, so that monetary and fiscal policies can be "fine-tuned" to the requirements of stabilization. Surprisingly, one of the major schools of thought on the matter, ***monetarism***, draws quite different conclusions.

Monetarism and monetary rules

Recall that the economists who make up the monetarist school attach limited importance to the multiplier process, and therefore to the ability of changes in the government budget to influence GNP. The reason for this attitude is their belief in the long-term stability of the velocity of money. They point to evidence like that presented in Figure 6. It shows a pronounced association between the M2 money supply and the money value of GNP. They argue that whenever GNP tends to change relative to the money supply, it sets off forces that bring it back into line. For example, an expansion of GNP caused by a rise in government expenditure runs up against a limited money supply. Interest rates rise. After a lag, higher interest rates reduce private investment, offsetting the expansionary effect of the government expenditure. This is ***crowding out.*** It also works in reverse. If GNP falls relative to the money supply because of a reduction in government demand, the interest rate drops and private expenditures rise to fill the gap.

The reverse of belief in the crowding-out phenomenon is a corresponding belief in the power of monetary changes. If GNP cannot for long fall or rise relative to a *fixed* money supply, then it must sooner or later follow a *changing* money supply. Control over the money supply, therefore, is control over the money value of GNP.

It is not, of course, control over real GNP. This is the message of the doctrine of the neutrality of money. If the money supply steadily grows faster than potential GNP, it must produce inflation equal to the difference between the two growth rates. Remember that persistent inflation is necessarily a monetary phenomenon. Monetarists are not the only ones who believe this. But monetarists also believe

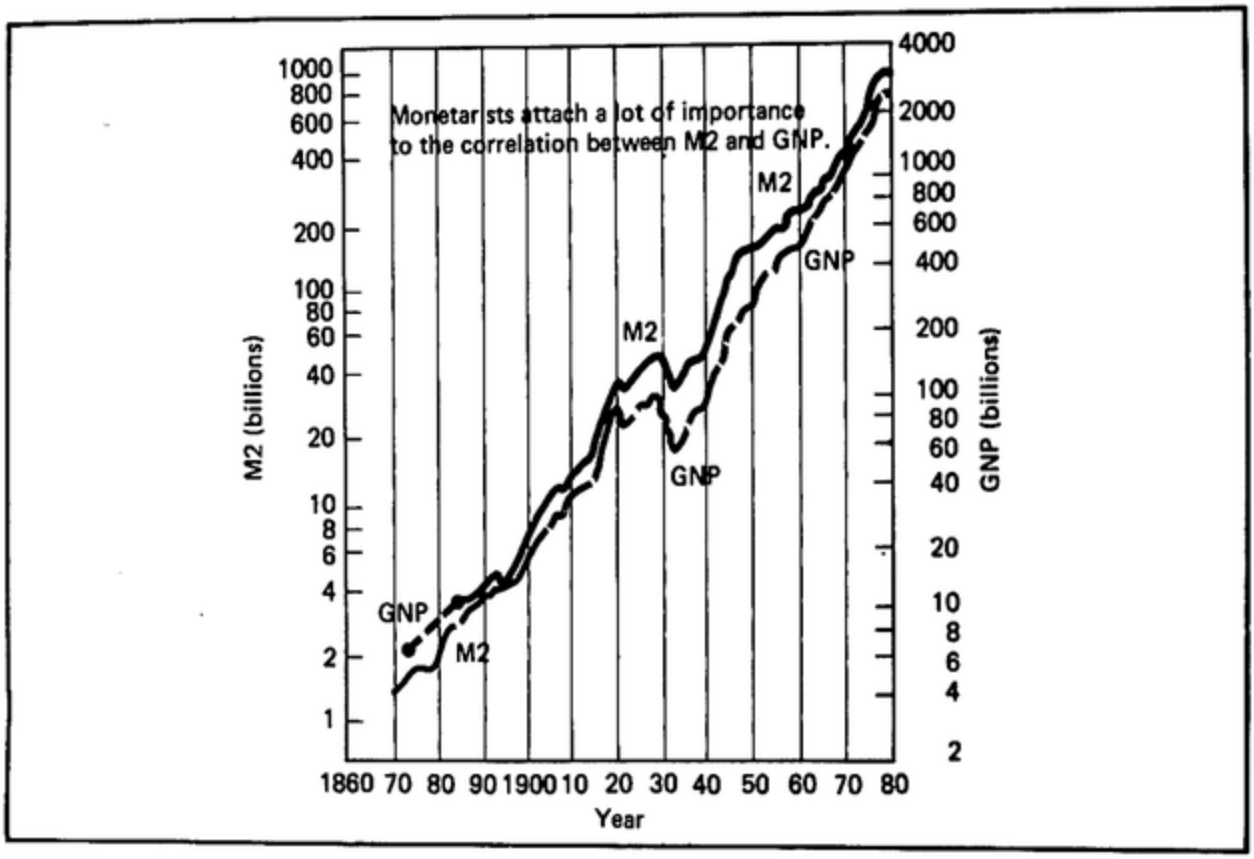


Figure 6 The M2 money supply and current-dollar GNP, 1870–1980
Source: *Economic Report of the President.*

that steady monetary growth at a rate smaller than that of potential GNP would result in steadily declining prices rather than falling output. Most of them have great faith in the flexibility of the price system and in the built-in stability of the market economy. They conclude that over the long haul, the economic system tends to maintain full employment and a rate of inflation that is directly determined by the growth of the nominal money supply relative to potential real GNP.

With all their emphasis on the short-run influence of money, you might suppose that monetarists would be vigorous advocates of stabilization through monetary policy. Yet one group of monetarists, led by Professor Friedman, is wholly opposed to *deliberate* attempts to stabilize GNP by changing the rate of growth of the money supply. The reason for this seeming paradox is Friedman's belief, based on a lifetime study of monetary economics, that there is a *long and variable lag*, as he puts it, between changes in the money supply and the ultimate consequences of those changes. If the Federal Reserve were to try to control a developing recession by rapidly expanding the money supply, it might be doing the right thing. But with equal or greater likelihood, it might find its policy taking effect much too late, adding momentum to an otherwise healthy recovery and carrying the economy on an excursion into inflationary terrain.

As a result of analysis of this sort, Professor Friedman has recommended that the Fed follow a rigid ***monetary rule***, deliberately restraining itself from trying to stabilize the business cycle. One such rule

would call for steady expansion in the nominal money supply at a rate of 3 or 3½ percent—the rate of growth of potential GNP. This particular strategy would lead to long-run price stability and would keep the Fed from causing short-run instability by ill-timed attempts at stabilization. Stabilization would be left to the private economy.

Monetarists also generally oppose using budgetary changes for stabilization. Although they think such changes are crowded out in the long run, monetarists do believe that fiscal changes have short-run effects. Because these effects are as likely to be ill timed as they are to be well timed, they may promote instability. The reason usually advanced for the poor timing of fiscal changes is not the same as that advanced for the poor timing of monetary policy, however. Monetary changes have a long lag in their effect on GNP, since they depend on changes in private investment plans. Such a delay in effectiveness is known as an ***outside lag,*** a lag in the public's response to the government policy. The lag usually attributed to fiscal policy is an ***inside lag.*** Because the institutions that control the budget are so cumbersome, the time lag between a cyclical downturn and the government's response is long and variable. Even though its impact may be immediate when it finally gets through the legislative process, the legislative process takes too long.

In many respects, the monetarist strategy for stabilization is a carefully thought out and documented version of the traditional distrust of government. Both monetary and fiscal policies are destabilizing in practice, it is argued. Greater stability could be achieved by getting the government out of the stabilization business. Since most monetarists think the government is far too large in most respects, they recommend dismantling many of its concrete spending programs and reducing taxes, shifting a substantial portion of control over resources from public to private hands. Carried far enough, this would greatly limit the government's ability to destabilize the economy. However, since monetarists assign such overwhelming importance to the supply of money, they do not recommend that monetary control be turned over to private hands. Reluctantly, one supposes, they wish to keep control of the money supply in the hands of the Fed. To ensure that this power is not misused, however, they propose to keep those hands tied by a rigid monetary rule.

Discretionary stabilization policy

Friedman's strategy for stabilization has never been tried in practice. All conscious attempts at stabilizing the economy have incorporated ***discretionary fiscal and monetary policies.*** Discretionary policy is the opposite of policy conducted by rigid rules. It involves perceiving a need to stabilize the private economy, formulating a policy to deal with the problem, taking action, and allowing the action to alter the course of economic events.

Potential GNP and the gap are extremely useful concepts for anyone who is trying to formulate discretionary stabilization policies. Potential GNP sets a target that is in the same units as the major performance indicator, actual GNP. The gap measures the magnitude of the GNP adjustment needed to reach the target. A Keynesian may use this gap, along with an estimate of the size of the multiplier, to gauge what government budgetary change is needed to reach the target. An advocate of discretionary monetary policy can make similar use of the gap—along with an estimate of velocity—to gauge what change in the money supply is needed to close the gap.

The next chapter will show you some of the ways in which discretionary policy is formulated and put into effect. The fact that we devote a chapter to discretionary policy should not be interpreted as an endorsement of this strategy or a rejection of the monetarist position. It merely recognizes how things are done. After you have seen how such policy is conceived and executed, you will be in a better position to decide for yourself whether it works well or badly.

Summary

This chapter has provided some of the background against which you should understand the government's attempts to stabilize output, employment, and prices. The most important things for you to remember are:

1. Because of the separation of responsibilities and powers within the U.S. government, it is difficult to achieve a coordinated stabilization policy.
2. The single agency most directly responsible for coordinating policy is the Council of Economic Advisers (CEA), which advises the President.
3. Most councils over the past two decades have argued that the government should seek steady growth along a path of GNP that results in a 4 or 5 percent unemployment rate. This path is known as potential GNP. If potential GNP were to be consistently maintained, the price level would remain fairly stable, barring such outside influences as the 1973 oil price increase or bad world harvests.
4. The cost of falling below potential GNP is measured in high unemployment and wasted output.
5. GNP can exceed potential, but this means venturing into the territory of the inflationary process. This can be done briefly with little lasting damage. But if GNP remains above potential for long, as it did during the Vietnam War, the stage is set for a persistent inflation.
6. The costs of inflation are mainly redistributive, shifting income and wealth around in a way that causes acute distress to some people.
7. If inflation proceeded at a steady rate that could be correctly anticipated, most of its redistributive effects could be eliminated.
8. Inflation brings on widespread fear and anger, even though its real costs are far smaller than those of depression. Thus, most democratic governments try to keep it within bounds.
9. It is costly to control a persistent inflation, since the most certain method of ending it is a deliberately engineered recession. This must be severe and long enough to break the expectations that sustain the inflation.
10. To keep GNP from straying very far from potential in either direction, many economists advocate a strategy of discretionary fiscal and monetary policies, designed to counter cyclical swings in the private economy.
11. One influential group of economists, the monetarists, doubts that discretionary policy can be timed well enough to do more good than harm. They think the economy has inherent self-stabilizing properties that can keep fluctuations within narrow limits, if the government would just stop making matters worse. Their program calls for small balanced budgets, steady moderate growth in the money supply, and general reliance

on private enterprise rather than government.

Key concepts

Stabilization policy
Council of Economic Advisers
Potential GNP
The gap
Hidden unemployment
Rational expectations
Indexing
Direct controls
Monetarism
Crowding out
Monetary rule
Inside and outside lags
Discretionary fiscal and monetary policies

Questions for review

1. Determine whether each of the following policy actions is the responsibility of the Office of Management and Budget (OMB), the Council of Economic Advisers (CEA), the President, Congress, Cabinet departments, or the Federal Reserve:
   a. Recommendation to the President on the choice of policies to control inflation or unemployment.
   b. Passage of the budget for the fiscal year.
   c. Projecting the costs of various spending programs.
   d. Presenting a budget to Congress.
   e. Administering federal programs.
   f. Identifying stabilization problems.
   g. Estimating budget surpluses and deficits.
   h. Setting targets for the rate of growth of the money supply.
   i. Planning a new federal program.
2. Why can the trend line for potential GNP be viewed as a low-inflation boundary?
3. a. Define GNP gap.
   b. What stabilization problem does a *negative* GNP gap indicate?
   c. What stabilization problem does a *positive* GNP gap indicate?
4. a. Define what is meant by *hidden unemployment.*
   b. What forms does this hidden unemployment take?
5. Explain why unanticipated inflation imposes higher costs than anticipated inflation.
6. a. Why is persistent inflation so much harder to end than temporary inflation?
   b. What are the costs of ending a persistent inflation?
   c. What are the costs of allowing persistent inflation to continue?
7. Are those who believe in the *rational expectations* school of thought optimistic or pessimistic in their view of the ease of ending inflation?
8. Suppose that you are a staff member of the CEA. Prepare a brief outline of the pros and cons of imposing wage and price controls as an anti-inflationary measure.
9. Monetarists believe that changes in the money supply have an important impact on the money value of GNP. Therefore, they are vigorous advocates of stabilization through monetary policy. True or false? Explain.
10. a. Define *crowding out.*
    b. Do monetarists believe that budgetary changes for stabilization will have a neutral impact on the economy? Explain.

31

Fiscal and Monetary Policy

As you read and study this chapter, you will learn:

- how to interpret the federal budget and its impact on total demand
- what the national debt is, who owns it, and what its significance is
- the various problems the Federal Reserve has in trying to control the money supply.
- some of the difficulties inherent in coordinating fiscal and monetary policies with each other and with other policies of the federal government

Some books take an awfully long time to get to the point, if they ever do. The sex manuals that were available when your parents were teenagers were about as much like *The Joy of Sex* as Lawrence Welk is like John Travolta. Often such a book was written by a clergyman, who would begin by telling you that sex is really OK and that you shouldn't feel guilty about being interested in it. The next two chapters would be sermons on the evils of promiscuity and disease. These would undo any reassurance gained from reading Chapter 1. After that, there would be chapters on anatomy and physiology, with diagrams about as explicit and helpful as the pictures in the Dick, Jane, and Spot books. Finally, the author would coyly pick his way through a chapter on how things are actually done. Presumably this is the part your parents were actually interested in. Picture their disappointment as the author hurried on to the final chapters extolling the joys of parenthood and family life. Picture them thumbing anxiously through the index, only to find that it skipped from "coition" to

"contrition" with no entry for contraception.

Don't waste a lot of time trying to figure out which of the earlier chapters in *this* book were secretly concerned with promiscuity and disease. Just be glad that you have finally reached the chapter that tells you how discretionary stabilization policy is actually carried out. It is divided into four parts, dealing with fiscal policy, the national debt, monetary policy, and policy coordination. The next chapter gives some case histories of how the principles of stabilization were applied during the 1960s, 1970s, and early 1980s.

The principles of fiscal policy

The preceding chapters have taught you a lot of macroeconomic theory. The experts responsible for framing stabilization policies use this theory when they decide what to propose to politicians. While economic analysis is often much more complex than anything in a text like this one, the actual theory used in making fiscal and monetary policy is readily understandable to anyone who has grasped the last nine chapters.

The federal budget

Fiscal policy is budget policy—taxing and spending. To think about it clearly, you must know what the federal budget looks like. Figure 1 shows in rough outline how it looked in 1981. As you can see, most federal revenue comes from personal income taxes and from contributions to various social insurance programs (actually taxes). Most of it is spent either on defense and related items or on transfer payments for health and income security. If you take the time to reflect on these magnitudes, the discussion of fiscal policy won't be as abstract as it might otherwise seem.

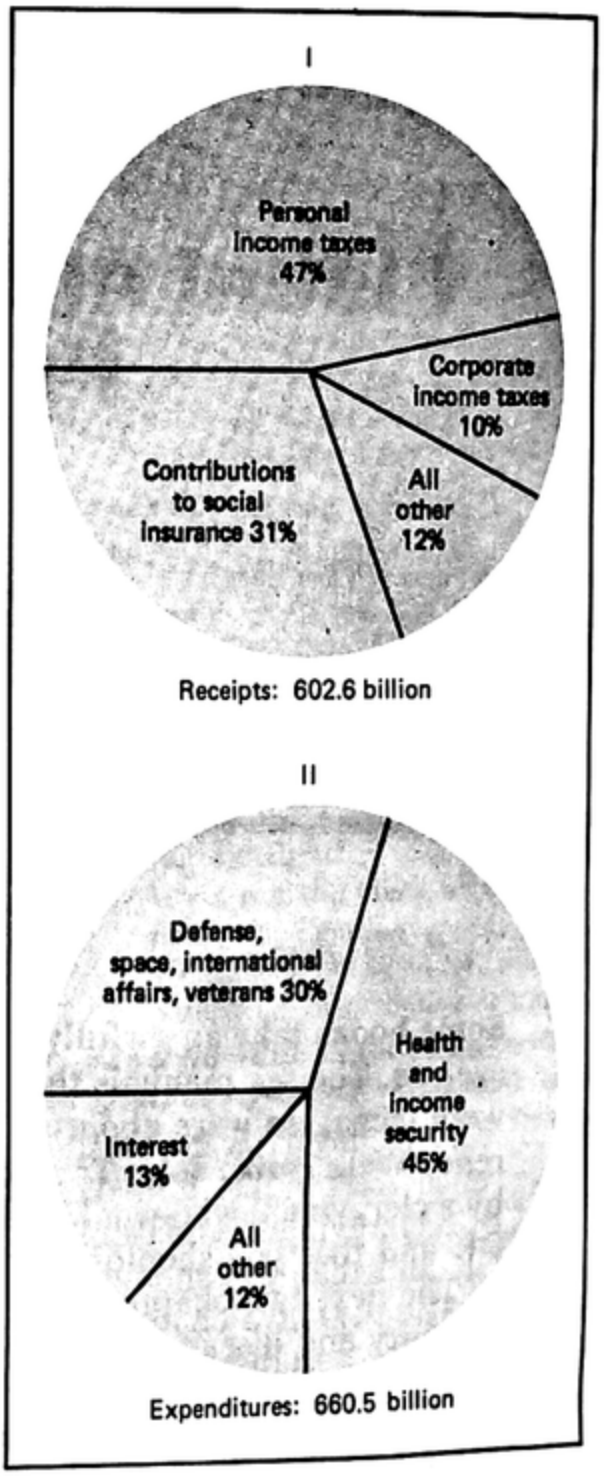


Figure 1 Distribution of receipts and expenditures in the federal administrative budget, fiscal year 1981

Most federal receipts come from personal income taxes and social insurance contributions. Most federal expenditures go for social programs and defense.

Source: *Economic Report of the President.*

In earlier chapters, we grouped budget figures into two broad categories, *purchases of goods and services* and *net taxes.* Government purchases are a direct demand for part of GNP. Net taxes, the difference between taxes and transfer payments, are a drain on the income of the private economy. Other things being equal, the larger net taxes are, the smaller private demands are. We will continue to group figures this way, but try also to think of them in terms of specific kinds of government expenditures and taxes. That will help you remember that the government does real things when it spends, and affects real people when it taxes.

The multiplier effects of changes in the federal budget

When you first encountered the multiplier, you learned to apply it to a change in the government budget. One of the things that makes the federal budget seem an obvious stabilization tool is the ability to get large GNP effects out of much smaller budget changes. Remember the multiplier theorem: An autonomous rise or fall in planned demand results in a multiplied effect on equilibrium GNP. Remember too, though, that when GNP goes up, taxes rise, and transfer payments automatically fall. These automatic budget changes help to stabilize GNP by making the multiplier smaller than it would otherwise be. Since these changes are already incorporated into the multiplier itself, their effects should not be counted a second time. All *built-in* or automatic changes in the budget are the result of a multiplier expansion or contraction, not the cause. The budgetary changes to which the multiplier applies are ***discretionary*** (deliberate) ***changes*** in purchases, transfer laws, and tax rates, which change the amounts of expenditures and net taxes at given levels of GNP. Concentrate on the deliberate-versus-automatic distinction. It will help you to sort out budget changes that trigger off the multiplier process from those that result from multiplier expansion and contraction.

The federal budget affects planned demand in two ways: First, an increase or decrease in federal purchases is itself an autonomous change in planned demand. Second, a cut or hike in net taxes collected at given levels of before-tax income results in an autonomous change in private planned demand. Either of these two changes triggers off the cumulative expansion or contraction we know as the multiplier process. When its force is spent, the autonomous demand change has been supplemented by an induced demand change, so that the overall rise or fall in GNP is some multiple of the autonomous change. Figure 2 will remind you of the arithmetic

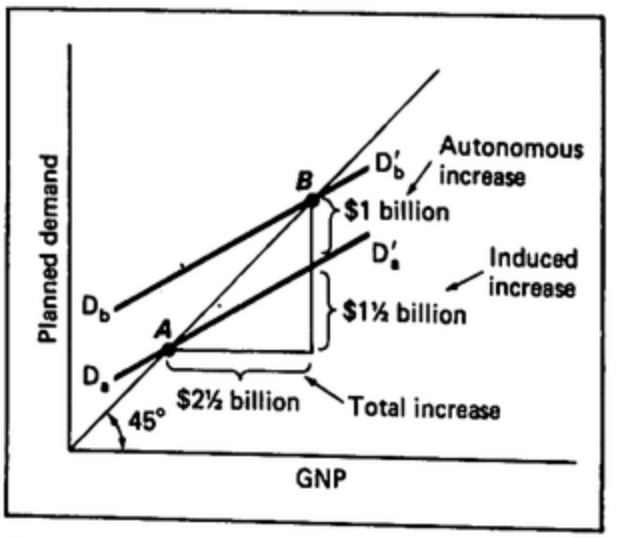


Figure 2 Multiplier effects of a $1 billion autonomous shift in planned demand, assuming a multiplier of 2½

A $1 billion autonomous rise in demand shifts the planned demand schedule from $D_aD'_a$ to $D_bD'_b$. The multiplier equilibrium moves from *A* to *B*. Since the slope of the planned demand schedule is .6 (as it is drawn), a $2½ billion increase in GNP will induce a $1½ billion rise in demand, so that the total rise in planned demand equals the rise in GNP.

behind this, something you have not seen for several chapters. In the figure, the multiplier number is 2½, which is roughly the right order of magnitude for the U.S. economy.

A deliberate change in net taxes is less powerful than an equal change in government purchases. Remember how tax changes affect planned demand. Suppose that income tax rates are cut, so that revenues drop by $1 billion. This leads to a $1 billion increase in disposable income. Will consumption rise by the full billion? No. Some of the higher after-tax income will be saved. Consumption will only rise by an amount equal to $1 billion times the marginal propensity to consume (MPC), which is less than 1. So a $1 billion tax reduction raises planned demand by less than $1 billion. A $1 billion increase in government purchases raises planned demand dollar for dollar. Both changes in planned demand are amplified by the same multiplier, but since the tax cut has the smaller initial impact, it also has the smaller overall effect.

Although there is an asymmetry between *purchases* and taxes, *transfers* and taxes have equivalent multiplier effects. That is the main reason for lumping them together as net taxes. Since net taxes affect demand only indirectly, by changing private income, some part of their impact is lost when saving changes. Government purchases are different from either taxes or transfers because they are themselves demand for GNP.

The balanced budget multiplier

A consequence of the difference between the multiplier effects of purchases and the multiplier effects of taxes is the ***balanced-budget multiplier*** theorem, which states that equal increases in purchases and taxes are expansionary on balance. Other things being equal, a larger balanced budget leads to greater planned demand than a smaller balanced budget does, even though both have a zero deficit. To see why, consider the following example. Suppose that the government raises its purchases of goods and services by $1 billion. This has two obvious effects: First, the government purchases *directly* increase demand by $1 billion. Second, they *indirectly* raise private demand by raising private income. If the government then increases taxes by $1 billion, this will only cancel out the indirect effect of the purchases increase, so that the combination of $1 billion in purchases and $1 billion in taxes will raise overall demand. Thus, a bigger federal budget expands total planned demand, even though the expansion in government spending is balanced by an equal increase in taxes. The same principle applies to state and local budgets. Although these governments do not conduct deliberate stabilization policy, their budgets do affect GNP. In 1929, federal, state, and local governments bought 8½ percent of GNP. In 1979, they bought 20 percent. If the level of government spending were cut back to 1929's relative scale, an enormous gap would open between potential and actual GNP, even if taxes were cut back to match.

Limitations on the multiplier effects of budget changes

In this discussion of budgetary changes, we have established three propositions about fiscal policy:

1. An increase or decrease in federal purchases has a multiplied effect on total demand for GNP.
2. An increase or decrease in net taxes makes total demand fall or rise by some multiple of the tax change, but the result of a tax change is smaller than that of an equal change in purchases.
3. A balanced change in purchases and

net taxes, either up or down, moves total demand in the same direction as the change in purchases.

If you have mastered the lessons of earlier chapters, though, you know that *any* multiplier analysis is subject to two major qualifications. First, the multiplier does not apply in the territory of the inflationary process. If resources are fully used, an increase in government purchases, for example, just creates excess demand. A contraction in government purchases may simply reduce excess demand without actually causing real GNP to drop. Second, the multiplier is always modified by the monetary feedback. A multiplier expansion causes interest rates to rise, and therefore crowds out some private demand. A multiplier contraction causes interest rates to drop, so that the effect of a decrease in government purchases is partly offset by rising private demand.

Because of this, it is not possible to analyze budgetary changes outside the context of actual conditions in the economy. If there is very high employment, a rise in government purchases will translate into pressure on prices and wages. Resources that are diverted to the production of goods for the government will not be available to produce goods for the private sector, and private demands will go unsatisfied. At very high levels of employment, there is a *trade-off* between production for the government and production of consumption and investment goods. But what happens during periods of high unemployment? Then, a rise in government purchases will put unemployed people to work producing goods for the government. Their incomes will rise, and they will be able to afford higher levels of consumption. If there are still unemployed people, this higher consumption demand can be satisfied, and production for the private sector will rise too, as an indirect result of rising production for the public sector. This is the kind of historical context in which things work out the way they do in Figure 2.

If government purchases expand in a context of great monetary "tightness," in which the Fed contracts the money supply to offset any increase in its velocity, or rate of turnover, the interest rate will rise to crowd out an equivalent amount of private expenditure. The multiplier process is impeded. But if the Fed "accommodates" the increase in government spending by expanding the money supply, it can keep the interest rate from rising. Then, the rise in government purchases will have its full multiplier effect, assuming that there are unemployed resources.

The built-in stabilizers

Net taxes absorb about 20 percent of any year-to-year change in gross national income and, because of how tax and transfer payment laws are written, change almost automatically in response to changing conditions. If you lose your job, your withholding and Social Security taxes stop along with your wages. It doesn't take an act of Congress to do it. If you qualify for unemployment compensation, you have only to stand in line to arrange to get your check, and in another line to arrange for food stamps, if you qualify for those too. Again, you don't have to wait for the Executive and Congress to take "discretionary action." There are no inside lags at all. Since they act so quickly, these ***built-in stabilizers*** limit the drop in your income even *before* it shows up in official statistics. If these stabilizing changes in the federal budget were not built into legislation already on the statute books, they would be caught up in the same political web that enmeshes so much of discretionary budget policy. The economy would be far more prone to cumulative instability than it is with a smaller multiplier.

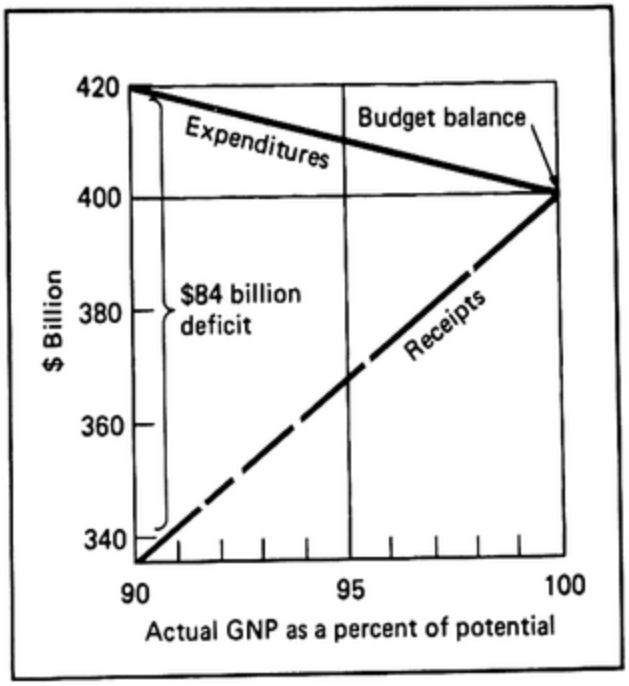


Figure 3 The built-in stabilizers

This diagram shows how the built-in stabilizers affect actual budget receipts and expenditures. As the expenditures and receipts schedules are drawn, both are equal to $400 billion when actual GNP equals potential GNP, so that the high-employment budget is balanced. But if actual GNP dropped to 90 percent of potential (with an unemployment rate of 9 or 10 percent), transfer payment increases would boost expenditures to $420 billion, and tax receipts would drop to $336 billion. The actual budget would show a deficit of $84 billion, even though there was no discretionary change in the budget.

Source: Authors' estimates, based on figures from *Economic Report of the President*. The scale is roughly correct for fiscal 1977.

To grasp the magnitude of the built-in stabilizers, look at Figure 3. It shows the receipts and expenditures in a budget that is balanced at $400 billion when GNP equals its potential value. But if GNP were only 90 percent of potential, this same budget would show a deficit of about $84 billion. At that level of GNP, federal expenditures would be about $420 billion rather than $400 billion because of higher unemployment compensation and welfare. Receipts would be about $336 billion rather than $400 billion, because lower incomes would produce lower tax collections. The entire difference would be due to the built-in stabilizers, not to any discretionary budget changes.

You must understand, however, that these built-in stabilizers cannot prevent fluctuations. They begin their work only when fluctuations actually occur. They are not a direct substitute for discretionary policy, which can be used to try to eliminate the business cycle rather than just to limit it.

Nonetheless, there are people who would like to leave the fiscal aspects of stabilization to the built-in stabilizers. You already know something about the monetarists and their skepticism of discretionary stabilization policy. They argue that the lags in the budgetary process make it nearly impossible to take deliberate fiscal action in a way that adds to stability. Since the economy has so much built-in stability, they think it is better to let the economy fluctuate within limits rather than to try, in an inherently clumsy way, to stabilize it further.

The high-employment budget

The changing seasons of the year make it hard to identify a "warm" day simply by its temperature. What is delightful in winter is frigid in summer. For similar reasons, the ups and downs of the business cycle make it hard to identify an "expansionary" federal budget simply by its deficit. A large deficit at full employment indicates a highly expansionary budget, but the same deficit in the midst of depression does not. The two deficits are interpreted differently because the magnitude of net taxes is partly determined by the prosperity of the private sector. When private demand is strong and the economy is prosperous, tax collections are high and transfers for unemployment compensation and relief are low. When the economy is depressed because private demand is weak, tax receipts are low and transfers are high. Thus, the size of the government surplus or deficit depends on the overall strength of private demand as well as on fiscal policy.

The Council of Economic Advisers has developed a method for keeping the bud-

getary changes that result from the business cycle separate from those that result from fiscal policy changes. This is the so-called ***high-employment budget.*** It is a valuable tool for charting changes in fiscal policy. The high-employment budget calculates receipts and expenditures that would result from existing legislation *if the economy were operating at its potential GNP.* Year-to-year variations in the high-employment budget reflect only three kinds of changes: (1) discretionary changes in expenditures (both purchases and transfers) that result from new expenditure legislation; (2) discretionary changes in tax receipts that result from new tax legislation; (3) changes in taxes and transfers that result from growth in population, potential output, and the price level. What do not appear in the high-employment budget are the automatic changes in the budget that occur when GNP departs from its potential. The built-in stabilizers are responsible for the discrepancies between the high-employment surplus or deficit and the actual surplus or deficit, which reflects the business cycle developments of the year to which it applies.

In Figure 3 we see how automatic changes in taxes and transfer payments affect the relative sizes of the actual and high-employment budgets. Since both expenditures and receipts schedules are drawn to equal $400 billion when actual GNP equals potential GNP, the high-employment budget is balanced at $400 billion. But if actual GNP dropped to 90 percent of potential (with an unemployment rate of 9 or 10 percent), transfer payment increases would boost expenditures to $420 billion, and tax receipts would drop to $336 billion. The actual budget would show a deficit of $84 billion, even though there was no discretionary change in the budget.

Figure 4 illustrates a different high-employment budget. It shows a $20 billion surplus, with receipts of $420 billion and expenditures of $400 billion when actual GNP equals potential GNP. If actual GNP equaled $400 billion, the actual budget would also show a surplus of $20 billion. If, instead, GNP were 90 percent of potential, the budget would show a $64 billion deficit.

It is important to understand that Figure 4 illustrates *only one high-employment budget*—with receipts of $420 billion and expenditures of $400 billion when actual GNP equals potential GNP. It illustrates *a range of possible actual budgets,* however, each corresponding to a different level of actual GNP.

Why might a government choose Figure 4's budget rather than Figure 3's? To see the answer, it is only necessary to note that Figure 4's budget is more restrictive or contractionary. Since it collects $20 bil-

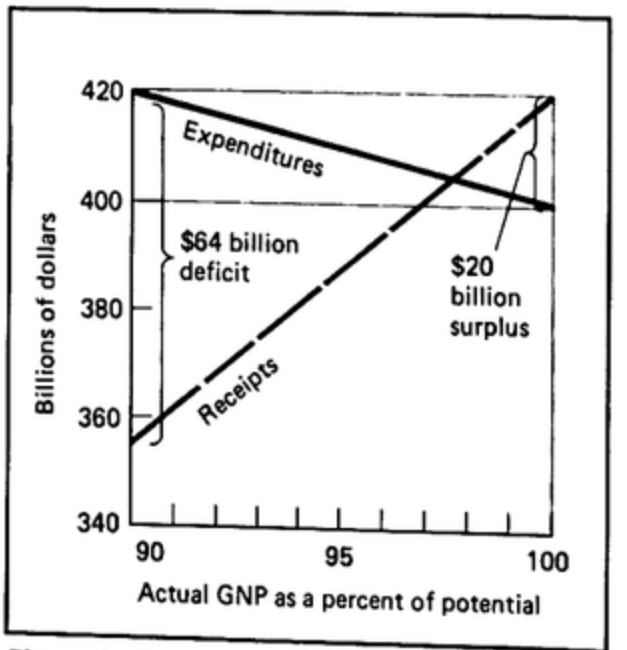


Figure 4 A more restrictive high-employment budget

This budget shows a $20 billion surplus at potential GNP, and greater receipts at every level of GNP than Figure 3's budget. It is, therefore, more restrictive than the budget depicted in Figure 3.

Source: Authors' estimates, based on figures from the *Economic Report of the President.* The scale is roughly correct for fiscal 1977.

lion more in taxes at every level of GNP, the economy will necessarily have much lower planned demand under Figure 4's budget than it will under Figure 3's. If the economy is operating in the territory of the multiplier process, it will have $40–$60 billion less real GNP under Figure 4's budget than under Figure 3's, depending on the precise size of the multiplier. If it is deep into the territory of the inflationary process, the higher taxes may not diminish real GNP very much, but they will reduce excess demand for both goods and labor, and moderate the rate of increase in wages and prices. You can see, then, that Figure 4's budget would clearly be more stabilizing than Figure 3's if the economy were operating beyond potential, with rising prices and the beginnings of a persistent inflation. The more restrictive budget would have been appropriate during the height of the Vietnam War, for example. Figure 3's budget would be preferred to Figure 4's in times of chronic unemployment, such as the early 1960s. Its greater expansionary impact would raise real income and employment relative to what would happen under Figure 4's more restrictive budget.

Many economists view the surplus or deficit in the high-employment budget as the best single measure of the stance of fiscal policy. They use the high-employment surplus or deficit relative to potential GNP to measure how restrictive or expansionary the budget is, and to compare alternative fiscal policies. If the high-employment budget changes toward greater surplus or smaller deficit, it has a contractionary multiplier impact on GNP: It is a restrictive change. If the budget moves toward greater deficit or smaller surplus, it is an expansionary change. The more substantial are expansionary and restrictive changes relative to potential GNP, the larger they are.

You can get some idea of the changes in fiscal policy during the 1970s from Figure 5. The only high-employment surplus in that period occurred in 1974. It was about ½ percent of potential GNP. The biggest high-employment deficits were in 1975 and 1977. They were about 1 percent of potential GNP. This narrow range of fluctuation in the high-employment budget reflects the "rock and a hard place" dilemma of policymaking during the 1970s. With both substantial unemployment and rapid inflation during most of the decade, policy advice was conflicting and ambivalent. It was hard to justify any vigorous action one way or the other. Between 1975 and 1978, budgetary policy was consistent from one year to the next.

You can also see in Figure 5 how misleading the actual budget can be as an indicator of policy. In 1974, the actual deficit increased, although the high-employment budget swung into surplus that year. In 1975, the actual budget showed an enormous deficit, nearly four times as large as

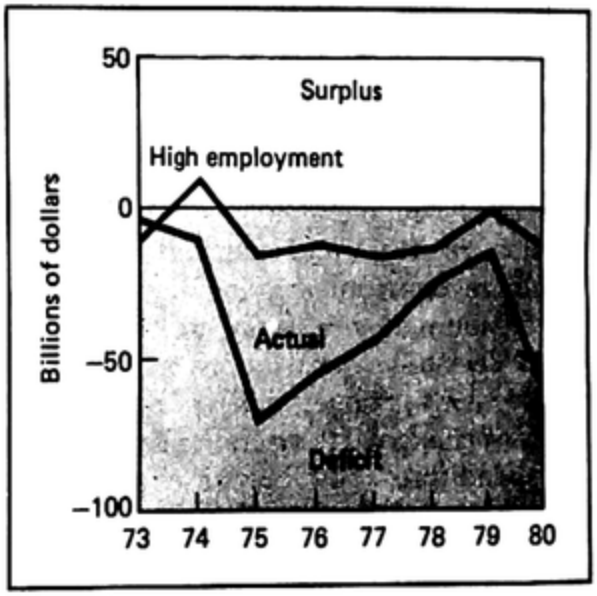


Figure 5 Actual and high-employment surpluses in the federal budget 1973–1980

In 1973, the GNP was close to its potential, and high-employment budgets showed a small deficit. Throughout the rest of the decade, when GNP was consistently below potential, the actual deficit was far greater than the deficit in the high-employment budget.

Source: *Economic Report of the President.*

the deficit in the high-employment budget. During 1976–1979, the actual deficit steadily dropped relative to the high-employment deficit. The reason for the differences between the two graphs is the action of the built-in stabilizers. Remember that in 1974 and 1975, the economy went into its worst slump since the 1930s. It gradually recovered over the following four years. Then it lapsed into another recession in 1980, and the actual deficit again got much larger than the high-employment deficit.

The national debt

The deficits in the federal budget mount up from year to year, except for the occasional year in which the budget shows a surplus. Their cumulative total is known, of course, as the ***national debt,*** something we all worry about from time to time. This worry is institutionalized in the form of a *statutory debt limit,* a lid on the size of the debt. Every time the debt limit is reached, the Congress holds hearings and reluctantly increases the limit. The whole thing is like trying to quit smoking by buying cigarettes a pack at a time. The result is not less smoking, but many more trips to the store.

The aspect of the debt that worries people most is its sheer size, over $1 trillion. They worry about national bankruptcy. People who suffer from this form of the Chicken Little Syndrome are particularly frightened because over a fourth of the debt matures every year. Isn't there some danger that the government won't be able to refinance this annual installment?

It would be a foolish economist indeed who would say in cold print that this could never happen. But consider three facts:

1. The financial community is full of institutions and individuals who are hooked on owning government debt. Government debt has a wide market among investors who show every sign of having confidence that the government's credit rating will always be the best. Long-term government securities always sell at yields below those on Aaa private bonds, a clear indication of their safety in the eyes of the market. Similarly, three- and six-month Treasury bill yields are always below those on prime commercial paper, the most comparable private issues. Remember that the lowest yields on securities of comparable maturity and tax status are always reserved for those issues with the lowest default risk.
2. When the federal government refinances its debts, it never finds itself in the position of New York City or Chrysler Corporation in the 1970s. Those institutions had trouble meeting their mature debt, simply because investors doubted their ability to meet interest payments on new debts incurred to pay off the old. With its immense taxing power and ability to cut other expenditures without going out of business, it seems incredible that the federal government would ever default on interest payments.
3. If the Treasury ever did suffer embarrassment in trying to sell new issues, the Federal Reserve would surely abandon its other stabilization efforts long enough to support the federal bond market.

If these facts don't make you more comfortable about the debt, look at Figure 6. As you can see, the expansion in the debt relative to the economy as a whole took place during the Great Depression and World War II. Compared to GNP, it is far less "oppressive" today than it was in 1945. Unless there is another Great Depression or World War, the debt is not

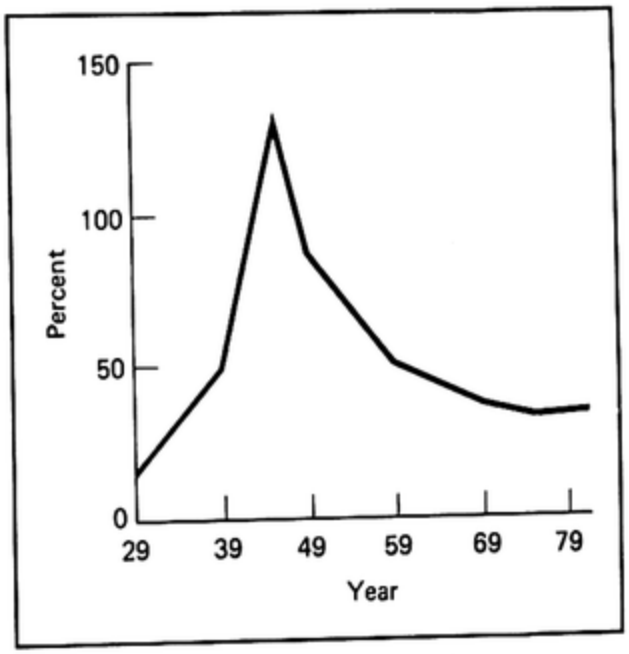


Figure 6 National debt measured as a percent of GNP 1929–1981

The debt relative to GNP peaked at the end of World War II and declined markedly until the early 1970s, when it leveled off. In 1981, the debt was about 35 percent as large as annual GNP.

Source: *Economic Report of the President.*

likely to increase much relative to GNP. And if the country faces another such catastrophe, the national debt will be the least of your worries.

Who owns the debt?

Still, the national debt is interesting. In particular, looking at the pattern of ownership of the debt will add to your understanding of financial institutions and practices. Figure 7 shows the distribution of the debt classified by owner as of September 1981.

Starting at about 2 o'clock on the circle and working counterclockwise, you can see that 9 percent of the debt was owned by state and local governments and 21 percent by federal agencies. Most of this represents reserves of government employee retirement plans and of the Social Security system.

Next is the 12 percent owned by the Federal Reserve. Remember its open market purchases of government securities? This portion of the debt represents the accumulation of past purchases. Since the Fed rebates nearly all of its interest receipts to the Treasury, this is hardly a debt at all, just a bookkeeping practice reflecting the legal autonomy of the Federal Reserve. If the Treasury itself performed the functions of the Fed, it would simply retire the securities it bought on the open market.

Now we come to the 44 percent owned by nonfinancial corporations, financial intermediaries, and individuals. These are assets of the private sector, held for private purposes. The holdings by nonfinancial corporations represent interest-bearing reserves of such firms, mostly held against future federal tax payments. The holdings of financial intermediaries represent a siz-

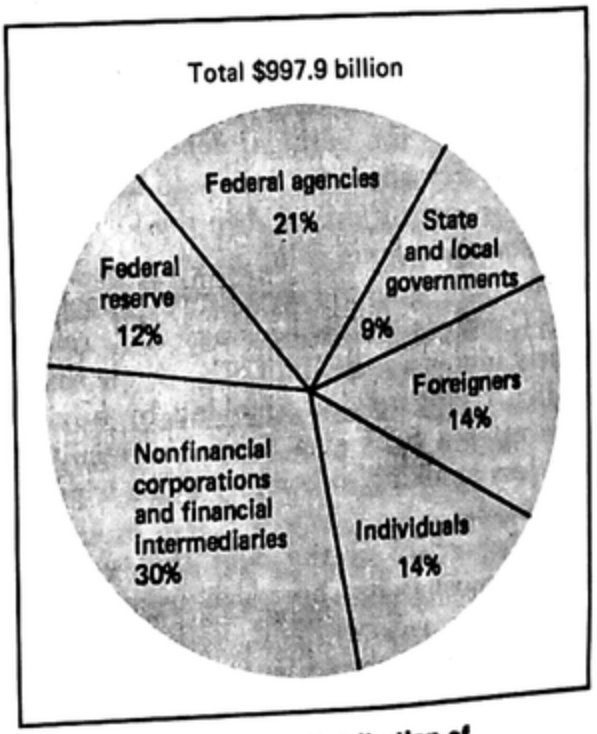


Figure 7 Percentage distribution of ownership of the national debt end of fiscal 1981

A third of the debt is owned by agencies of the federal government itself. Another sizable portion is owned by state and local governments. A small portion is owned by nonfinancial corporations. Nearly a quarter belongs to financial intermediaries and indirectly to their depositors. Individuals own directly about 14 percent of the debt, mainly in the form of savings bonds. About a sixth is owned by foreigners (this share has grown in recent years).

Source: *Federal Reserve Bulletin.*

able portion of the earning assets of banks, thrift institutions, mutual funds, life insurance companies, and pension funds. Indirectly, these assets are owned by depositors, who receive much of the interest. In many ways, they resemble the part of the debt owned by individuals. By putting their funds into financial intermediaries, individuals gain access to the interest on large-denomination federal securities. Most of the debt that is owned directly by individuals consists of small-denomination, low-interest savings bonds.

Many people are bothered by the redistributive aspect of the national debt. No matter what high-employment budget the government chooses, it must collect more taxes to make higher interest payments. Neither these taxes nor the interest is reflected in *net taxes*, from which interest is deducted. But the taxes come from the average taxpayer, and the interest goes to the owners of the debt. In times when high interest rates are coupled with legal ceilings on the rates payable by banks and thrift institutions, much of the interest goes to people who can afford to buy large-denomination securities or certificates of deposit instead of having to put their money into a savings account. This looks very much like taxing ordinary people for the benefit of the wealthy. To a considerable extent, it is. It would be far more egalitarian if we had some high-interest federal debt available in small denominations. We do not, largely because the government is unwilling to compete directly with privately owned banks for savings. Although the debt itself is not an instrument for redistributing income to the rich, the way it is denominated has this effect when interest rates are high.

We have not yet mentioned the final slice of Figure 7's pie, the 14 percent of the debt owed to foreigners. This is a genuine debt, largely due to the oil price increases of the 1970s. The oil-producing countries permitted us to run a trade deficit during this period, accepting federal (and private) debt in exchange. If they had not, we would have had to pay for our oil with increased exports, and our level of domestic consumption and investment would have been lower as a consequence. Higher exports would have crowded them out. If we have to pay off this debt in the future, we will incur this crowding out then. If you want to worry about the federal debt, this part of it is a good place to start.

The burden of debt

Seeing the difference between the foreign and domestic portions of the debt should help you distinguish between truths and fallacies about the debt. The domestically owned part of the debt is not a "burden on our grandchildren." If it is ever paid off, Peter's grandchildren will be burdened to pay Paul's, and a certain amount of taxing Peter to pay Paul will always be involved in paying interest. For the domestic portion of the debt, the burden is redistributive only. Paying interest and principal on the foreign portion, however, will someday involve taxing Peter and Paul to pay Ali. Well, Ali's grandparents are selling us oil on credit, aren't they? If they weren't, we would be paying now rather than later.

To see where the real ***burden of the*** domestically owned ***debt*** lies, look at the time when the debt is incurred. Suppose that the government increases its purchases during a deep depression. If it puts the expenditures into public works, cultural enrichment, or labor force training programs, it will benefit the country in the future. If it cuts taxes to encourage the private sector to undertake similar programs of investment in physical capital or human resources, the future of the country will also be the better for it. When there are many unemployed resources, there will be no crowding out. Indeed, the multiplier ef-

fects of such policy will probably induce additional capital formation and human development. Our future selves and our offspring should all welcome a debt so productively incurred.

Suppose, however, that the government runs a deficit at full employment. It crowds out an equivalent amount of private expenditure. How should you, as a representative of future generations, view this deficit? You should look at the benefit to you of the least valuable government project and compare it to the benefit (again, to you) of the private resource uses crowded out. If you see the benefits as greater than the opportunity costs, you should welcome the debt. If the balance of benefits and costs goes the other way, you should see the debt as a symbol of a real burden.

Of course, the last paragraph is pure social science fiction of a very fantastic sort. People don't think this way, don't want to think this way, and couldn't if they did. And you wouldn't want your children to marry such people, either. But like a lot of fiction, it contains a genuine commentary on life. It pinpoints the real issues surrounding the debt. Deficits that are incurred during depressions doubtless benefit the future. Deficits that are incurred during prosperity probably do not, since the *marginal expenditures* (those most easily done without) are often of little benefit to anyone, living or yet unborn. This doesn't mean that we should never undertake government expenditures at full employment, only that we should clearly recognize that they have a cost to people coming after us as well as to ourselves. Nor should you think that this cost applies only to public expenditures. Private consumption also crowds out investment at full employment. Whenever we consume rather than invest for the future, we "burden our grandchildren."

The conduct of monetary policy

Monetary policy is conducted by the Federal Reserve, mainly through its control over the reserve position of its member banks. Much of what the Federal Reserve does is intended to help accommodate the banking system, not to whip it into line. The Fed was originally established to provide the country with a "flexible currency," a money supply that would respond to the changing needs of business. At the time, it was well understood that flexibility must be kept within limits, particularly on the long-run growth in the money supply. But it was not envisioned that the Federal Reserve would try to stabilize the business cycle. This is a fairly recent role, one the Fed has partly chosen for itself and partly had forced upon it. Monetary stabilization policy is, therefore, conducted by means that were designed for other purposes. That accounts for its puzzling indirectness.

Fiscal policy involves spending and taxing in big, blunt ways to affect the circular flow of goods and services. There is nothing subtle about it. But when the Fed buys government securities to increase bank reserves to encourage lending to increase the supply of money and credit to lower interest rates to encourage investment, it sounds more like "The House That Jack Built" than a serious stabilization measure.

The differences between monetary and fiscal policies sometimes lead people to imagine that they have different targets. Since fiscal measures affect the circular flow directly, it is tempting to see them as controlling production and employment. Because money supply growth is so closely tied to persistent inflation, it is easy to imagine that monetary policy is exclusively aimed at the price level, rather than at output and employment. That is not so,

and neither monetarists nor Keynesians think it is so. **Changes in the nominal money supply affect both output and prices by changing the demand for goods, the same path that fiscal policy treads. That means, of course, that fiscal changes also affect both output and prices. The money supply's special position in stabilization comes from its decisive role in determining the *long-term* trend in prices.** This role makes the money supply particularly newsworthy in times of inflation, especially when an episode of inflation threatens to turn into a chronic problem. To the Keynesians, the long-run link between money and prices is the main reason for taking an interest in monetary policy. In the short run, however, they view monetary policy as simply a flexible but limited tool, to be used in cooperation with fiscal policy. The monetarists' position is not much different in kind, save that they play up the short-run influence of money on both output and prices, and downplay the importance of fiscal policy.

You are already familiar from earlier chapters with the Fed's means of controlling the money supply. The next section is mainly for review, therefore.

Open market policy, reserve requirements, and discounting

The Fed has three major tools it can use to regulate the rate of growth of the money supply. The first of these is its authority to buy and sell government securities on the open market. You remember how ***open market operations*** work from our analysis of banks and deposit creation in an earlier chapter. If the Fed buys securities, it pays for them with a check on itself. When the check is deposited in the seller's bank and presented to the Fed for payment, it is credited to the bank's reserve account. This adds to the lending capacity of the banking system. If, instead, the Fed sells securities, it debits the buyer's check against his or her bank's reserve account. This reduces the banking system's lending capacity, forcing a contraction of loans. Expansion and contraction of bank lending is simultaneously an expansion and contraction of the money supply. Banks create deposits when they make loans, and extinguish deposits when loans are repaid.

As you can see from Figure 8, the rate at which the Fed expands bank reserves through its purchases of government securities varies a lot from year to year. Far from following a rigid monetary rule, it conducts its open market policies very deliberately. The Fed has often been criticized for the inconsistencies in its policies and for their seeming irrelevance to the economic problems of the times. We will talk more about that in the next chapter.

A second lever by which the Fed controls banks is its power to change ***reserve requirements.*** Remember that banks must keep reserve deposits and vault cash in amounts that depend on the size of their

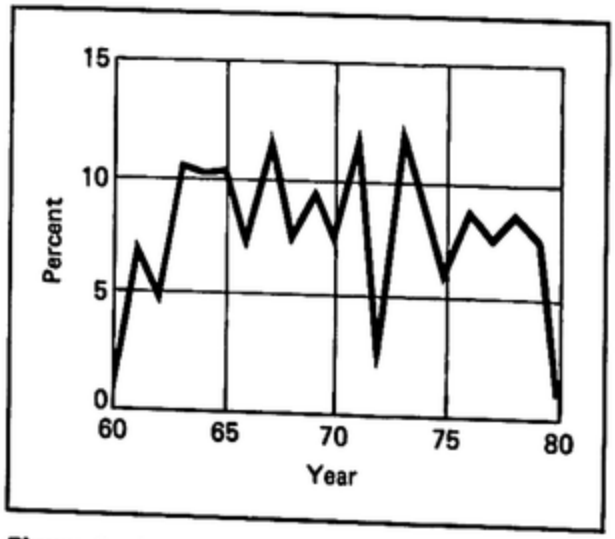


Figure 8 Annual percentage changes in the Federal Reserve's holdings of U.S. government securities 1960–1980

The Fed's pattern of open market purchases is very erratic from year to year.

Source: *Economic Report of the President.*

deposits and the distribution of these deposits among demand, savings, and time categories. If the Fed raises required reserve ratios, it forces banks to reduce their lending to build up reserves. If it lowers reserve ratios, the Fed creates lending capacity by freeing reserves that had been held to meet requirements.

Although the Fed is constantly using open market operations to change the actual reserves of the banking system, it seldom changes reserve requirements. Since reserve requirement changes affect all banks at the same time, they are very powerful and are generally used only when drastic action is called for. Under the circumstances, the power to change the reserve requirement is not usually a very helpful tool.

However, the *existence* of reserve requirements is important in the functioning of monetary policy. Think of the reserve requirement as the fixed jaw of a vise or clamp. When open market operations change the actual reserve position of the banks, their lending capacity depends on the gap between the changing reserves and the required reserves as they are fixed by the volume of deposits. If banks were completely free to choose their reserve holdings, they could move their target reserves up and down as the Fed changed actual reserves. This would weaken the effectiveness of open market operations.

A third policy tool is the Fed's control over the interest rate it charges banks that borrow reserves. Banks that face reserve shortages may borrow at the Fed to cover their shortfall. This gives them "breathing space" to contract their loans and gradually build up reserves. The ***discount rate*** is the rate that the Fed charges those banks that avail themselves of the opportunity to borrow reserves. When the Fed sets the discount rate below the rate on *federal funds*, which are reserves loaned from one bank to another, it in effect encourages

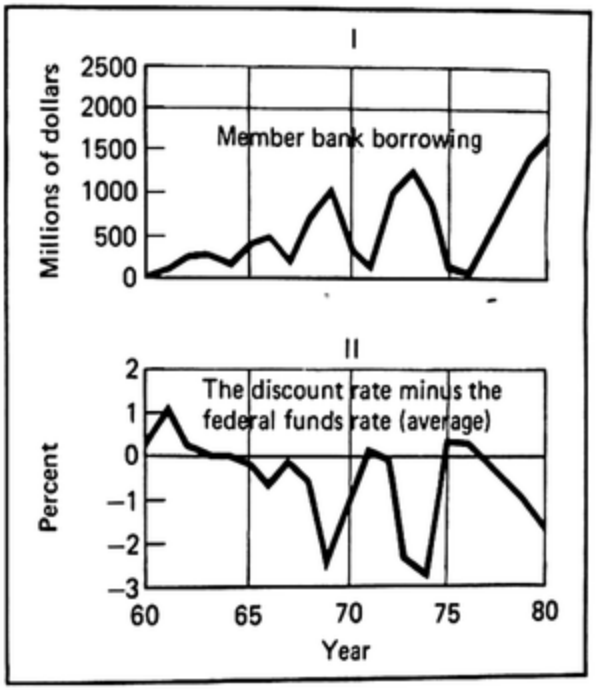


Figure 9 Member bank borrowings from the Fed and the discount rate minus the federal funds rate 1960–1980

Member bank borrowings are usually largest when the discount rate is a bargain.

Source: *Economic Report of the President.*

banks to borrow reserves. As you can see from Figure 9, banks borrow a lot from the Fed when the discount rate is a bargain, and very little when it is not.

By setting a low discount rate, the Fed "goes easy" on banks with a reserve shortage. It does so when it wishes to avoid forced loan contraction. When it sets the discount rate high, it puts greater pressure on banks that are short of reserves. In effect, variations in the discount rate soften or harden the reserve requirements by making it cheap or costly to borrow to meet reserve requirements.

Control over the money supply

Through its open market operations, reserve requirements, and discount policy, the Federal Reserve operates on the reserve position of the banking system. This is not the same thing as controlling the

money supply. For several reasons, there is a lot of slippage between what the Fed does and what happens to the money supply. Open market operations and discount policy control the ***monetary base:*** the sum of currency and reserve deposits. Reserve requirements set outer limits on the money that may be created from a given monetary base. But within those limits, the banks and the public will have a lot to say about the size of the money supply.

Each of the following actions on your part will tend to increase the money supply relative to the monetary base:

1. You minimize your currency holding, keeping all your money in the bank. When you hold currency, it subtracts dollar for dollar from bank *reserves.* Put it in the bank, and the banking system can multiply each dollar of reserves several fold through multiple deposit creation.
2. You minimize your checking account, keeping most of your money in savings deposits, on which there are no reserve requirements.
3. You do your banking at a small bank whose required reserve ratio is low.

Your bank also can do its part by never keeping excess reserves. Either it is always fully loaned up in the service of its customers, or it lends any excess reserves on the federal funds market.

If you and your bank behave this way, no one will accuse the two of you of originality. You will simply be following the trends of the past few decades. Figure 10 shows the result: a steady uptrend in the ratio of M2 to the monetary base. If the Fed wants to control the money supply, it cannot count on constancy of this ratio to help it meet its targets. It must carefully monitor what is happening to the money supply and compensate for the offsetting behavior of the banks and the public. This

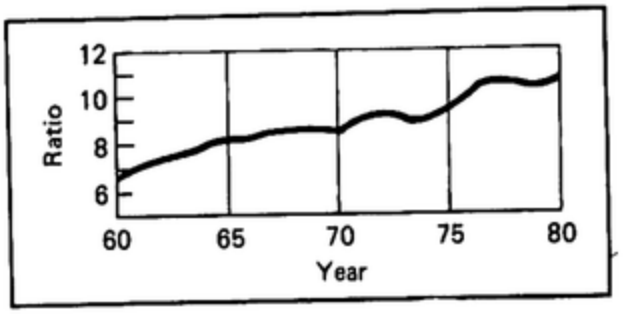


Figure 10 The ratio of M2 to the monetary base 1960–1980

One of the problems of monetary control during the 1960s and 1970s was an erratic upward trend in the ratio of M2 to the monetary base.

is not easy. The Fed has not always succeeded in hitting its targets, even though they are fairly wide ranges rather than pinpoint growth rates. For example, the Fed announced in the fourth quarter of 1975 a growth target for M1 of $4^1/_2$ to $7^1/_2$ percent over the coming year. In fact, it came fairly close to the middle of this range for the year as a whole. But for the first four months, M1 grew at a rate far below $4^1/_2$ percent a year. The reason for this failure was not simple incompetence, but rather the unpredictable slippages between changes in the monetary base and changes in the money supply itself. It is very hard for the Fed to separate trends in the relationship between the money supply and the monetary base from erratic month-to-month changes that have little significance. Control of the money supply is much harder in practice than it looks at first glance.

Control over interest rates

If the Federal Reserve tries specifically to control interest rates, as it frequently has in the past, its job is even more complicated than if it just sets a target for the money supply. Even if it achieves steady growth in the money supply, it will not achieve stable interest rates, which depend on the demand for money as well as on its

supply. Moderate growth in the supply of money in the face of briskly growing demand means excess demand for credit and rising interest rates. The same moderate growth in the supply of money in the face of sagging demand means falling interest rates. Even if the Fed followed a rigid monetary rule, there would still be cyclical fluctuations in interest rates because of the GNP effect built into the monetary feedback. **Cyclical variability in interest rates is one of the built-in stabilizing properties of the economy. It helps to stabilize investment through the interest effect of the monetary feedback, much as the fiscal stabilizers help to stabilize consumption.**

This built-in tendency for interest rates to rise and fall with the cycle has caused many people to urge the Fed not to use the interest rate as an indicator of how "tight" or "easy" its monetary policy is. The most persuasive voice has been that of Professor Friedman, who has for years argued that the only true gauge of monetary tightness is the money supply itself. His argument has particular force in times of inflation. The interest rate peaks of the 1970s were extraordinarily high by comparison with the past. This suggests the severest monetary tightness. Yet in those times of double-digit inflation, the *real* interest rate was not very high. Sometimes it was negative. So was money tight or easy? Friedman's answer is that you should look at the rate of growth of the money supply. During 1974, for example, consumer prices rose more than 12 percent. The M2 money supply grew about 7 percent. The real money supply, therefore, fell, since prices grew faster than the nominal stock of money. By the money supply criterion, money was tight.

In 1976, M2 grew at a rate of 11 percent, well above the 7 percent rate in 1974. Since the inflation rate was *lower* in 1976, the difference in *real* rates of monetary growth was even more pronounced. Money was unquestionably easier in 1976 than in 1974, by Friedman's measure.

You don't have to swallow Friedman whole to see the wisdom in what he has to say about this. Because of the close long-run relationship between the money supply and GNP, it looks as though money supply growth will *eventually* influence prices, output, or both. But if you believe in Friedman's "long and variable lag" in the economy's reaction to changes in the money supply, you should not expect to see its influence right away. Nor should you expect to be able to read it *in any simple way* from changes in interest rates. It follows that the Fed ought not to expect this either and therefore ought *not* to judge its monetary policy by looking at what is currently happening to interest rates. In fact, the Federal Reserve itself seems to be coming around to Friedman's view.

Policy coordination and conflict

It almost goes without saying that fiscal and monetary policies should be coordinated with each other. It would be tragicomical if the Fed were aggressively expanding the money supply to fight unemployment while Congress and the Executive were raising taxes to fight inflation. It is also fairly evident that the specific programs funded by the budget ought not to work against the needs of stabilization. Yet, ***policy coordination*** is not always easy, and policymakers have sometimes seemed to be working at cross-purposes.

Most of these policy conflicts can be traced to one or more of the following circumstances:

1. The various policy planners cannot agree about what should be done, and end up at actual cross-purposes, sim-

ply because there is no obvious right thing to do.

2. The planners agree and are working in concert, but a combination of inside and outside lags gets their policy effects badly out of synchronization with each other and with the needs of stabilization.
3. The planners agree about the needs of stabilization, but are constrained in their behavior because the government does a wide range of things, many of which are inconsistent with the needs to stabilize output, employment, and prices.

The first of these policy conflicts stems from a combination of faulty or inconclusive information and the separation of powers. Such conflicts could be resolved if we understood the economy more perfectly, so that it would be easy for different agencies to reach the same conclusions about the facts. Or they could be resolved if the decision-making structure were more unified. But unity without understanding could make things worse. The Fed, Executive, and Congress might all agree to go in the same wrong direction.

The second of these three policy conflicts is the sort of Keystone Kops scenario that makes Professor Friedman grimace instead of laugh. The problem of timing is the basis for his skepticism about discretionary stabilization policy. Again, better information and more streamlined policy-making procedures might shorten the lags and better coordinate the policies, assuring that they work in the same direction at the same time. One hopes they would work in the right direction.

The third kind of policy conflict results from genuine dilemmas that cannot be resolved by greater intelligence and cooperation. The high-employment, price-stability trade-off is a dilemma: One good thing pitted against another. When the Fed fights inflation by tightening credit and flattening the home-building industry, it is trading one person's good against another's. The building industry isn't usually very grateful. Much of the agony and contradiction in stabilization policy stems from conflicts among goals. When the government spends and taxes, real people are affected. A steam turbine presumably does not care that its speed is stabilized by a feedback device. But people care deeply about taxes, interest rates, and government projects. Many of them are themselves powerful, are represented by powerful lobbies, or are the constituents of powerful legislators. Unlike the steam turbine, they fight back. Even if they did not, the government would be sensitive to needs other than those of stabilization alone.

Part of understanding economic policy is seeing how it fits into the broader context of economic, political, and social life. To see how stabilization policy is limited by this broader context, look at the following examples.

National defense spending

The abstract theory of stabilization policy usually treats government expenditures as one of the tools of stabilization. But thinking this way can be a serious mistake. Look at Figure 11, for instance. The top line, which traces the fever chart of wars both hot and cold, shows federal defense purchases as a share of potential GNP. The extreme swings in defense spending dwarf those in any other major component of demand. Since they aren't offset by contrary swings in federal civilian expenditures, they are a large part of the instability problem.

This may give you a headache. Aren't government purchases one of the tools of stabilization policy? How can the solution

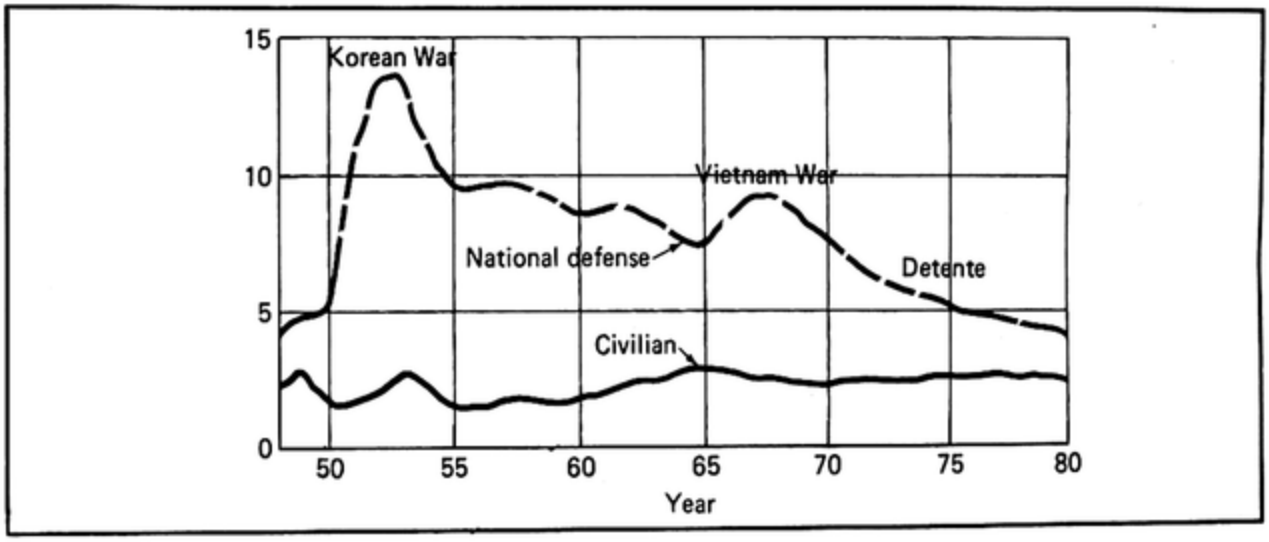


Figure 11 Federal purchases of goods and services as a percent of potential GNP 1948–1980

Federal purchases for national defense have been a major source of instability since World War II. Purchases for other purposes have been much smaller and far more stable. Their fluctuations, though minor, have generally been timed to contribute to stability.

Source: Authors' calculations, based on *Economic Report of the President.*

to instability be part of the problem? Remember that at the beginning of the chapter we warned you to think about the budget in terms of specific programs to keep from thinking about stabilization in too abstract a way. The defense budget may be affected in small ways by current stabilization problems. But in its big swings, it is governed by current tensions and struggles between capitalism and communism, between American military power and that of the Soviet Union, China, Vietnam, and North Korea.

In principle, the government could raise at least enough taxes to offset the expansionary effects of higher military purchases, or even more if stabilization seemed to demand it. In practice, it was unable to do so during either the Korean or Vietnam wars. For example, from 1951 to 1953, the unemployment rate hovered around 3 percent. But in both 1952 and 1953, the federal budget had substantial deficits even though GNP was well beyond potential. During the Vietnam War years of 1966–1969, the unemployment rate was under 4 percent for four years in a row, yet only in 1969 did the budget show a surplus. In 1968, with an unemployment rate of 3.6 percent, the federal national income budget had the largest deficit it had shown since World War II. The President and Congress were not facing up to the need for higher taxes to prevent inflation when GNP was beyond potential.

During both of these wartime episodes, interest rates were raised considerably above previously normal levels. By the interest rate criterion, the Fed was fulfilling its responsibility for stabilization. By the criterion of monetary growth, it did well during the Korean War, though less so during the Vietnam War period. The real villains, however, were the Executive and Congress, who did not raise the taxes needed for noninflationary war finance. Because concerted stabilization policy would have called for much higher taxes than the political process could impose for wars that did not directly threaten the na-

tion's security, fiscal policy worked at cross-purposes to the needs of stabilization.

Financing transfer payments

Another, somewhat subtler example of policy conflict can be seen in the history of transfer payments and contributions for social insurance. As you can tell from Figure 12, transfers have become a far more important part of personal income than they were 25 or 30 years ago. In 1953, federal transfers were only about 5 percent as large as wages and salaries. By 1980, the ratio of transfers to income from work had risen to 18 percent. Part of the increase can be attributed to a rise in the number of retired people relative to those working. Part is due to more adequate survivor and retiree benefits under Social Security. And part is due to newer programs such as Aid to Dependent Children (ADC), Medicare, and Medicaid. The most rapid increase began in the mid-1960s, roughly coincident with the beginning of what developed into the persistent inflation of the 1970s. Obviously, increases in transfer payments had to be offset by increases in tax rates. Otherwise, the budget would have swung toward massive high-employment deficits.

Transfer payments in this country have traditionally been financed by taxes on wages and salaries. This is particularly true of Social Security, which was originally a contribution-based retirement plan. As you know, the benefits have not always been limited to contributions plus accrued interest, but there remains a strong presumption in favor of raising *payroll taxes* (that is, contributions to social insurance) whenever Social Security benefits are increased. You can see from Figure 12 that the increase in social insurance contributions has therefore been substantial over the period of transfer increase, al-

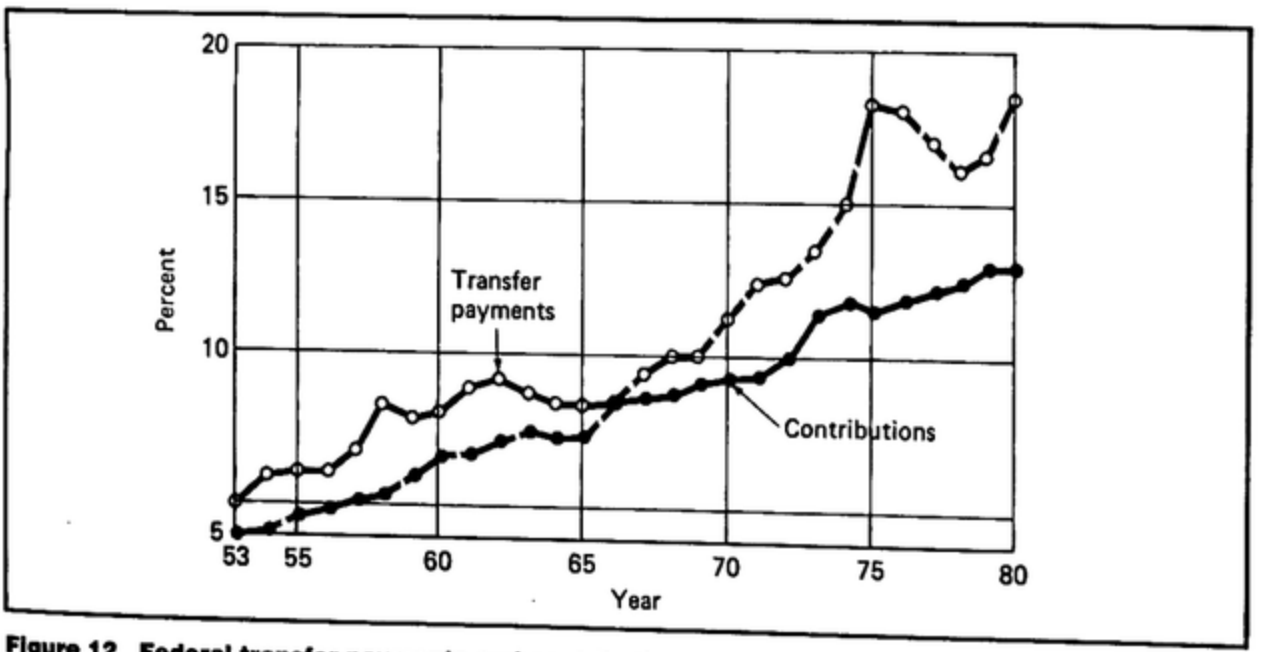


Figure 12 Federal transfer payments and contributions to social insurance as a percent of wages and salaries 1953–1980

When social insurance contributions are raised to offset rising transfer payments, they tax away purchasing power, but they also raise labor costs and prices.

Source: *Economic Report of the President.*

though it has not quite kept pace with expenditures.

The fact that payroll taxes offset much of the transfer increase during the 1950s, 1960s, and 1970s made the budget more restrictive. These taxes took income out of private hands and restrained private demand. This kept the unemployment rate higher, and wage and price increases lower, than they could have been. Particularly in the late 1960s and 1970s, these tax increases helped to stabilize the price level at a time when it was rising rapidly.

Remember, however, that permanent increases in labor costs are eventually reflected in prices, unless they are offset by productivity gains. And what are labor costs made up of? Wages, fringe benefits, and payroll taxes. The increases in payroll taxes linked to rising Social Security benefits directly affected costs and prices. Between 1965 and 1979, federal social insurance contributions rose from 7 percent of wages and salaries to 14 percent, about ½ percent per year. If this had all been "passed forward" as higher prices, it would have added about ½ percentage point to the annual rate of inflation. Doubtless some of it was "passed back" as lower money wages, but a strong case can be made that most of it was passed forward. Because the government was constrained by political tradition to finance higher transfers through a payroll tax rather than simply to raise income taxes by an equivalent amount, its policy was contradictory. An income tax would have taken away real purchasing power by lowering money income relative to prices. The payroll tax took it away by raising prices relative to money income.

Numerous other federal programs contributed to the inflation of the 1960s and 1970s, complicating rather than complementing the policies directed at controlling prices. Among them were various regulations on employers designed to improve product safety, make the workplace safer and healthier, and preserve the environment. Many such programs were beneficial and long overdue. But all of them imposed costs on business firms and raised prices to some degree in an era of serious inflation. We point this out not to impose conservative values on you—you can make up your own minds on such normative issues—but to make you realize, as a matter of fact rather than value, that many of these specific programs ran counter to the main thrust of stabilization policy.

Summary

This chapter has been concerned with the nuts and bolts of fiscal and monetary policies. The major points that you will want to remember are the following:

1. Autonomous changes in federal purchases and net taxes have multiplier impacts on GNP.
2. Because part of the impact of taxes falls on savings, tax changes affect GNP less, dollar for dollar, than changes in government purchases. Hence, even a balanced change in the budget has a multiplier effect.
3. Because taxes and transfers change automatically as private incomes change, the structure of tax and transfer laws is built into the multiplier process. Because they cushion disposable income against changes in GNP, taxes and transfers are known as built-in stabilizers.
4. Like all multiplier reactions, the effects of budget changes are conditioned by the availability of unemployed resources and by the reactions of the credit markets.
5. The built-in stabilizers cause federal receipts to rise and fall with the busi-

ness cycle. Thus, the surpluses and deficits in the actual budget are very misleading indicators of fiscal policy. A far better indicator is the high-employment budget, in which receipts and expenditures are presented as they would be if the economy were operating at potential GNP. If the high-employment budget moves toward surplus, it is getting more restrictive. If it swings toward deficit, it is getting more stimulating.

6. The national debt results from past deficits in the federal budget. Despite many people's worries about it, the government securities that represent the debt always command an eager market. The debt has been falling in relative size since the end of World War II. Nearly all of it is owned internally. Only the part owned by foreigners is in any real sense a "burden" on future generations. However, deficits that crowd out investment at full employment do impose a cost on the future economy.
7. Monetary policy, like fiscal policy, operates to stabilize the economy through affecting aggregate demand. Fiscal policy affects the circular flow quite directly. Monetary policy affects it indirectly, through the credit markets.
8. The Federal Reserve conducts monetary policy mainly by controlling the reserve position of its member banks. The principal tools by which it does this are its open market purchases and sales of government securities, its control over the required reserve ratios of its members, and its ability to change the discount rate at which it will lend reserves to the banking system.
9. The ability of the Fed to control the money supply is limited by the reactions of banks and the public to changes in the reserves it provides. Its control over interest rates is further limited by public reactions to changes in the money supply.
10. Although it is obviously desirable that fiscal and monetary policies be coordinated with each other and with the needs of stabilization, this is not always possible.
11. Failures of policy coordination usually result from disagreements about the right policy, lags in the implementation and effectiveness of policies, and conflicts between stabilization requirements and other goals of government policy.

Key concepts

Discretionary changes
Balanced budget multiplier
Built-in stabilizers
High-employment budget
National debt
Burden of the debt
Open market operations
Reserve requirements
Discount rate
Monetary base
Changes in the money supply versus changes in interest rates

Questions for review

1. a. Define built-in stabilizer.
   b. Does the presence of built-in stabilizers make the multiplier larger or smaller? Explain.
   c. Will any change in tax revenues cause the multiplier process to begin? Why or why not?

2. Suppose that the following statement was excerpted from the speech of a local politician: ". . . And while my proposed program will cost $1 billion, it need not have an inflationary impact on demand if it is coupled with a $1 billion increase in tax revenues." Is the statement true or false? Explain.
3. You are asked to determine whether fiscal policy was more expansionary under the Carter or the Reagan administration. You begin by looking at the actual surpluses/deficits over the four years of each administration. Is this information sufficient to indicate the tightness or ease of fiscal policy? Is other information needed? Explain.
4. True or False: The national debt
   a. is so large that there is real danger of national bankruptcy.
   b. is a burden to our grandchildren.
   c. involves a redistribution of income from the average citizen to the wealthy.
   d. held by foreigners is more of a burden than the domestically held portion.
   e. will involve a lower burden if it is incurred at a time of high unemployment.

   Discuss.
5. a. Define monetary base.
   b. What actions by the public and the banking system will cause a given monetary base to support a smaller money supply?
6. "The Federal Reserve engages in monetary policy by influencing the rate of growth of the money supply. Yet none of the tools of monetary policy used by the Fed directly influences the money supply." Explain.
7. "Cyclical fluctuations in the interest rate help to stabilize the economy." Explain.

32

Stabilization Policy: The Historical Record

As you read and study this chapter, you will learn:

- how the administration, the Congress, and the Federal Reserve tried to cope with problems of stabilization during the 1960s, 1970s, and early 1980s
- how their efforts were complicated by the Vietnam War, the food shortage of the mid-1970s, and the OPEC oil cartel
- that there was a fairly consistent thread running through the policies of both Democratic and Republican administrations until Reagan's presidency
- that stabilization efforts were sometimes successful and sometimes not
- that there are some general lessons to draw from these experiences

Picture yourself at the wheel of one of those giant hook-and-ladder fire trucks that have a second steering wheel at the back. You are hurtling down the Oregon coast on Highway 101, in one of its famous morning fogs, with a stuck throttle and a loose steering linkage. The person in command (?) of the back steering wheel has a hangover. Worst of all, your mother and father are in the cab with you, telling you how to drive. As the ultimate hairpin looms out of the fog, you see that all is lost. With a giant effort you reach for the engine bell, to warn everyone that you cannot stop. As it clangs out, it gradually turns into your alarm clock. You jump from bed, sweating profusely, and run to the window. There, in front of your house, is the White House limousine, waiting to take you to your job as chairman of the Council of Economic Advisers. You wish you were back in your fire engine.

Actually, things are not always this bad. During recent decades, there have been instances in which stabilization policy worked well, the way its designers had expected it to work. There

have been other times when it was possible to conceive of an effective policy, although political and administrative crosscurrents made it impossible to execute it. But there have also been genuinely nightmarish situations, in which the economy was in such an inherently contradictory position that is was hard to put together a policy that would straighten it out painlessly.

This chapter surveys some of the highlights of the stabilization experience during the 1960s, 1970s, and early 1980s. Now that you have worked your way through the past 10 chapters, you should be able to see many of these historical developments as part of an integrated pattern. You know some of the facts of recent economic history. More important, you now know a lot of macroeconomic theory. As we survey the past, you should be able to blend fact and theory into an integrated account of what happened. As you do, you will gain the confidence to think independently about the economic events of your own adult lifetime. These events are not "elemental," like tornadoes and sunny afternoons. They are the working out of the logic of distinctively human institutions that often work well on their own, sometimes can be made to work better, and sometimes fall apart despite our best—or because of our worst—efforts.

The early 1960s

Let us start with a pleasant story about an episode in policymaking that worked out well and seemed for a while to promise a "new era" in the stabilization of the business cycle. This was the period of analysis, debate, and consensus that produced the Revenue Act of 1964, a substantial cut in income taxes specifically designed to close a persistent positive gap between actual and potential GNP.

Symptoms: Weak recoveries from the recessions of 1958 and 1961

From all outward appearances, the recession of 1958 was a perfectly ordinary cyclical downswing. The mid-1950s had been exceptionally prosperous, with unemployment around 4 percent for three years in a row. The boom was sustained by a high rate of investment in inventories and equipment, along with its multiplier effects. After this period of accumulation had run its course, however, there followed a sharp decline in the rate of investment, which fell by 15 percent in real terms from 1956 to 1958. A sharp drop in investment is almost inevitable after a period in which capacity and inventories grow much faster than GNP. Once firms have an adequate stock of capital, they have no immediate incentive to undertake further accumulation. While the built-in stabilizers kept consumption from actually dropping, the plunge in investment was sharp enough to cause a very slight dip in overall GNP. Since potential GNP was steadily rising, a GNP gap opened up, and the unemployment rate in 1958 reached 6.8 percent, up $2\frac{1}{2}$ points from 1957.

In 1959, GNP rose by 6 percent, as investment rebounded sharply. This lowered the unemployment rate to $5\frac{1}{2}$ percent. But then the recovery fizzled. In each of the next two years, actual GNP growth was only about $2\frac{1}{2}$ percent, not enough to keep up with growth in potential GNP. As a result, the unemployment rate rose. By 1961, it was back up to 6.7 percent. Never between 1958 and 1961 did the unemployment rate drop to the 4 percent level typical of the mid-1950s.

Recovery from the 1961 recession at first seemed normal. GNP in 1962 was about 6 percent above its 1961 level, and the unemployment rate dropped to $5\frac{1}{2}$ percent, in what looked like an instant replay of the 1959 recovery. Unfortunately, this recovery fizzled in much the same way

its predecessor had. The unemployment rate bottomed out at 5.3 percent in July 1962 and then crept upward. By the beginning of 1963, it was up around 6 percent again. Business fixed investment in 1963 was only 14 percent higher than it had been in 1957. A 2 percent annual rate of increase in fixed investment is not enough in itself to keep planned demand growing at the same rate as potential GNP. Clearly, something was chronically wrong with the economy.

Diagnosis: Business investment and the federal budget

When John F. Kennedy took over the White House in 1961, he brought into the Council of Economic Advisers a group of economists with strong Keynesian convictions. Because of their Keynesian training, they generally viewed the American economy as subject to bouts of instability. In their view, however, it could be stabilized by an active program of discretionary policy, with particular emphasis on fiscal policy.

Since they believed so firmly in the need for discretionary fiscal policy, one of their first programs was aimed at increasing its effectiveness. In January 1962, the President asked Congress for "standby authority" to cut taxes temporarily during a recession (subject to congressional veto) and to initiate at his discretion a limited program of public construction. This authority would have substantially reduced the "inside lag" between the time when the need for discretionary action is recognized by an administration and the time when such action is taken. As you might expect, Congress viewed these proposals with considerable hostility. If enacted, they would have materially changed the balance of power between the President and Congress. It therefore declined to grant him these discretionary powers.

At the same time, the administration sought to achieve better coordination between monetary and fiscal policies. The President asked the Congress to make the term of office of the chairman and members of the Board of Governors of the Federal Reserve System coincide with that of the President himself. This would make it possible for a new president to pick a board with economic policy views similar to those of his administration. Again, the Congress declined to strengthen the presidency.

Besides proposing to increase the flexibility and coordination of stabilization policy, the Kennedy Council analyzed the causes of the incomplete recovery from the recessions of 1958 and 1961 and concluded that the underlying incentives for private investment were too weak to keep demand rising in step with potential GNP, given the degree of restriction built into the high-employment budget. In effect, there was a structural inconsistency between the federal budget and the behavior of the rest of the economy that was causing the period of rapid multiplier expansion during recovery to falter before GNP reached its potential level.

The Council's analysis was based on a variant of the kind of multiplier theory with which you are familiar. Remember the following propositions about the multiplier process:

1. Whenever planned demand differs from production (GNP), production moves toward planned demand, rising when demand is bigger than GNP and falling when it is smaller.
2. When planned demand and production are equal, the multiplier process is in equilibrium.

Now look at how this applies to the relationship between the federal budget and the rest of the economy. For simplicity,

suppose that the rest of the economy consists only of private firms and households, with no foreign trade or state and local governments. In this case, the condition of multiplier equilibrium is:

$$C + I + G = GNP = C + S + TN$$

(consumption plus planned investment plus government purchases equals consumption plus total saving plus net taxes). Now rearrange this to read:

$$TN - G = I - S.$$

This says that the multiplier process is in equilibrium when the federal sector surplus *(TN − G)* equals the planned private sector deficit *(I − S)*. The private deficit represents an excess of planned demand relative to production; the federal surplus represents a shortfall. They must balance each other in multiplier equilibrium.

Sometimes a recession originates in the public sector. This invariably happens at the end of a war. The federal high-employment budget swings toward surplus, triggering off a multiplier contraction. Usually, however, recessions originate in the private sector. When they do, investment nearly always leads the decline. Businesses decide that further expansion in capacity and inventories will not be profitable. They cut back planned investment, throwing the circular flow out of equilibrium. GNP drops. If it were not for the built-in stabilizers, GNP would have to fall far enough to bring the private deficit back into balance with an unchanged federal surplus. Saving would have to drop as far as investment. But because of the stabilizers, the federal budget moves toward deficit in recession, balancing off some of the drop in investment and limiting the decline in GNP.

Think about what happens in recovery. After inventories are worked off and capacity has stopped growing for a while, a resumption of capital growth seems profitable to business firms. Planned investment rises, leading the recovery in a process that simply reverses the sequence of events of the recession. The private sector swings toward deficit, and the multiplier expansion in GNP must carry far enough so that the rises in private saving and the federal surplus just balance the rise in planned investment.

Now concentrate on the role of the built-in stabilizers. They are responsible for the rise in the federal surplus as GNP goes up. As you know, this makes the multiplier smaller than it would otherwise be. A given autonomous rise in investment leads to a smaller GNP increase than there would be without the built-in stabilizers. Is this good or bad for an economy whose government is trying to achieve high employment and price stability?

That depends on the size of the GNP gap relative to the increase in investment. If the gap is large relative to the investment increase, then the built-in stabilizers keep the recovery from carrying GNP to its potential level. If the gap is small relative to the investment increase, the stabilizers keep the expansion from carrying *too far*, past potential GNP and into inflationary territory. Ideally, the federal surplus should rise to balance the private deficit just at potential GNP, neither short of nor beyond it.

In applying this analysis to the concrete problems of the late 1950s and early 1960s, the Council pointed out that the mid-1950s had seen very intense capital investment. Much investment in plant and equipment is based on profitability calculations over a horizon measured in years or even decades. Following a period of relatively high investment in expansion and modernization, business will not immediately embark on another such program until it has "grown into" the large modern facilities it has just built. It would not be profitable to do so. Because of this, it is not

surprising that investment was weak following the boom of the 1950s. In 1963, for example, investment in plant and equipment was only 11 percent higher in real terms than it had been in the peak year of 1957. Yet potential GNP had grown 23 percent over the same period.

When investment incentives are weak, private investment will be small relative to saving at potential GNP. If the federal budget has built into its tax and expenditure provisions a large high-employment surplus, that budget is structurally incompatible with a small private sector deficit. The result is a chronic inability of the economy to reach full employment because GNP reaches equilibrium well below potential. Compatibility requires that the high-employment surplus just match the private deficit at *potential* GNP.

The prescription: A tax cut

Once the Council thought it understood what was wrong, its next step was to propose a remedy that would bring the budget into line with the realities of private investment. The 1963 *Report* proposed a general reduction in tax rates, so that taxes would siphon off less private purchasing power.

The President asked for a reduction in personal income taxes of $8 billion a year and a cut in corporate taxes of $2.5 billion a year. This amounted to about 1.8 percent of the 1963 GNP. The intent was to shift the planned demand schedule upward, raising private purchasing power at every given level of GNP. Allowing for a tax multiplier of 2, this would have raised the level of actual GNP relative to potential GNP by about 3.6 percent. Since each percentage point rise in GNP relative to potential lowers the unemployment rate by about 0.4 percentage points, the effect on the unemployment rate would be a decrease of about 1.4 percentage points, nearly enough to get down to a 4 percent unemployment rate from the 5.6 percent ruling in early 1963.

Despite vigorous lobbying by the Council and other agencies, a doubtful Congress failed to act on the tax cut in 1963. It was reluctant to cut taxes when the budget was already in deficit. Only in 1964, after Kennedy had been assassinated and Johnson had become President, was a tax reduction passed. What followed was a period of rapid expansion in output and employment. By the end of 1965, the unemployment rate was down to 4 percent. Everything seemed to have worked out just as the policy planners said it would.

You will notice that the Council's analysis relied on strictly Keynesian arguments. The Kennedy Council always gave lip service to monetary policy, but its heart belonged to the budget. Walter Heller, Kennedy's Council chairman, wrote a few years later that he thought of the period as "the completion of the Keynesian Revolution." By this he meant that these were the first years in which public policy was specifically guided by Keynesian principles. (Public policy has yet to be guided by monetarist principles in any consistent way. This is one reason the controversy continues. The two horses have never run the same course.)

The air in Washington was thick with the aroma of self-congratulation. The Johnson economic team had won the Big Game! In fairness, it should be pointed out that by March 1966, Heller was already worrying in public about the economic implications of rising expenditures for the Vietnam War. But there were no monetarists in sight to point out that the money supply (M2) had grown at a rate of 7 percent in 1964 and nearly 9 percent in 1965, and that this alone might account for the rapidity of growth in GNP. For the moment, it was the Age of the Economist or, more accurately, the Age of the Keynesian Economist.

This is a good time to stop for a moment and reflect on something: Just because the people making policy at that time thought that they lived in a Keynesian economy doesn't mean that the world was following a Keynesian script. History is not always governed by the books its actors learned their lines from. You have probably often seen a TV program, movie, or play with someone, and then argued with that person over what was *really* going on. Suppose you had a chance to argue with the actors themselves. Who is to say that their interpretation of their own roles would necessarily be the last word on what actually happened on the stage?

The late 1950s and early 1960s would not embarrass a monetarist. He or she would point to the slowdown in monetary growth from 1956 to 1957 followed by the recession of 1958, the speedup in monetary growth from 1957 to 1958 followed by the recovery, the subsequent slowing down of money growth from 1958 to 1960 leading to the 1961 recession, and the resumption of more rapid money growth leading to the 1962 recovery. All of this can be seen in Figure 1. The relatively high unemployment rates during the recoveries would cause no embarrassment either. A monetarist might well ask, "What is so magical about 4 percent unemployment? Weren't prices starting to rise uncomfortably fast in the mid 1950s? Isn't 5½ percent unemployment much more compatible with maintenance of price stability?"

Whatever the facts of the matter, stabilization policy in the early 1960s was formulated by Keynesians, who prescribed Keynesian medicine.

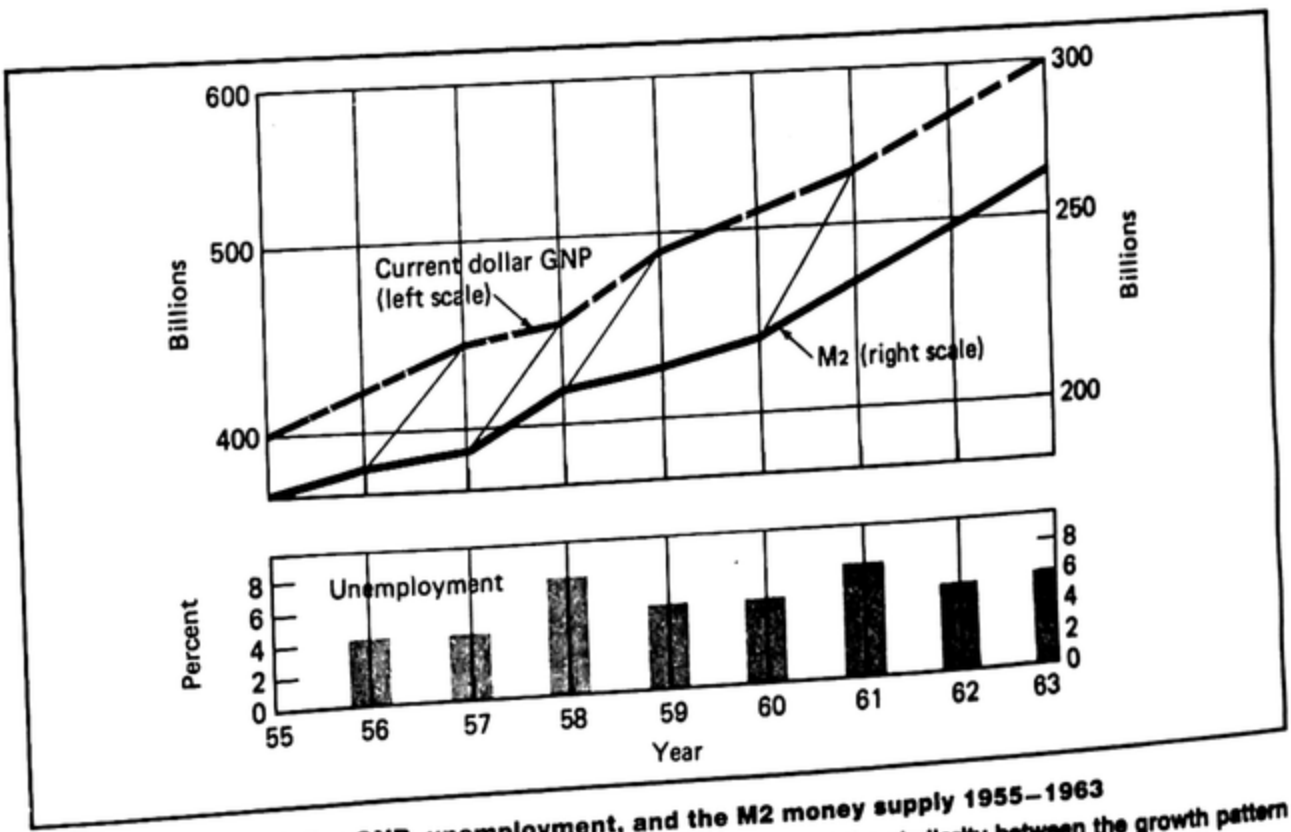


Figure 1 Current-dollar GNP, unemployment, and the M2 money supply 1955–1963
A monetarist analyzing the late 1950s and early 1960s would point out the similarity between the growth pattern of M2 and the subsequent year's growth patterns of GNP.
Source: *Economic Report of the President.*

The late 1960s

The Age of the Economist soon gave way to the age of the B52 bomber. The Vietnam War probably produced more civil strife than any episode in our history since the Civil War, more even than the Great Depression. And Walter Heller's fears about its impact on the economy were fully justified.

You have already learned about the relationship between the war expenditures and the development of persistent inflation. There is no point in going over that ground again. But there are two instances of attempts to stabilize the economy that you can learn from—the ***credit crunch*** of 1966 and the ***tax surcharge*** of 1968.

The credit crunch of 1966

As the economic response to the 1964 tax cut gathered momentum, so did the war. Official government estimates show Vietnam expenditures rising from under $1 billion in fiscal 1965 to a peak of almost $29 billion four years later. The high-employment surplus fell by about $15 billion between mid-1965 and the end of 1967. Relative to the size of the economy, this was the fiscal equivalent of the 1964 tax cut all over again. Not many people's personalities are markedly improved by a second martini.

The administration and Congress were in a very uncomfortable position. They had just passed the tax cut, which had been politically popular. Now, as the price level started to rise more rapidly, the stabilization case for a general tax increase grew stronger. But the war was increasingly unpopular, and the political climate increasingly hostile. The economic situation called for a tax increase. The political situation vetoed such a move. Under these conditions, it was impossible for the government to put together a fiscal policy that was both politically and economically stabilizing.

An additional bind on fiscal policy came from Johnson's "War on Poverty," a variety of programs designed to improve the lives of people at the bottom end of the income distribution. Since the Johnson administration had made a strong political commitment to improve the health, education, job skills, neighborhoods, and transfer incomes of the poor, it had little room to cut back civilian expenditures to make space for the growth in military spending. This is a very good example of what we were talking about at the end of the last chapter. Since the expenditure side of the budget is made up of real programs, not abstract dollars, it is not very flexible. This hampers fiscal policy a great deal.

With fiscal policy in a double bind, the entire burden of stabilization fell on the Federal Reserve and monetary policy. While the fiscal side of government continually ran a multibillion-dollar deficit at full employment, the Fed single-handedly tried to cool the economy down. Even though the rate of inflation during most of 1965 was not very alarming, the Fed correctly concluded that the rapid growth in GNP would soon lead into inflationary territory. It acted accordingly. Toward the end of the year, it raised the discount rate to its highest level since 1930. In mid-1966, it raised reserve requirements by a full percentage point, a very large increase relative to the reserve ratios on most deposits. Throughout 1966, it engineered a considerable slowdown in the rate of growth of the money supply, coordinating its open market policy with the changes in the discount rate and reserve requirement.

Some of the results of Fed policy show up in Figure 2. Relative to current-dollar GNP, the growth in M2 faltered in early 1966. During midyear, monetary growth

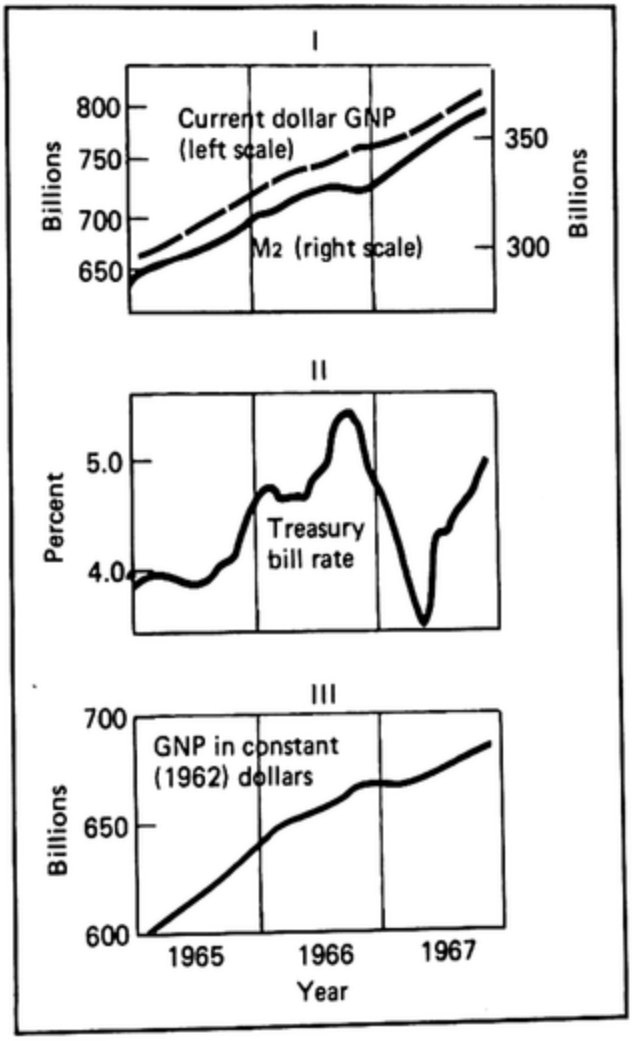


Figure 2 The credit crunch

During 1966, the rate of growth of the money supply tapered off sharply relative to the rate of growth of current-dollar GNP. There followed a sharp rise in interest rates, a slowdown in the rate of growth of real GNP, and finally, a slight drop in real GNP in the first quarter of 1967.

Source: *Economic Report of the President.*

actually stopped altogether. At a time when current-dollar GNP was growing at a rate of nearly 8 percent a year, this forced an enormous speedup in velocity. The rate of interest on Treasury bills reached its highest level since 1921. Real investment dropped by 16 percent between the fourth quarter of 1966 and the second quarter of 1967. As a consequence, real GNP itself dropped slightly from the fourth quarter to the first quarter and rose only slightly in the second quarter. This pause in expansion was generally known as a "minirecession." It did not register at all in the unemployment rate, but there was a definite decline in overtime and in the length of the average work week. The rate of increase in both wholesale and retail prices dropped off considerably as goods markets temporarily slackened.

The most important part of this episode is what it shows about the power of money. If you are like most students (and quite a few professional economists), you probably find the theory of fiscal policy much more convincing than the theory of monetary policy. The budget affects the demand for goods and services in very direct, almost self-evident ways. But monetary policy is so indirect that discussions of how it works always sound a little suspicious, maybe even dishonest. Well, look at what happened in 1966 and 1967. And remember that during this period federal purchases were growing at an annual rate of over 15 percent in real terms. Can you seriously doubt that what happened was the effect of the credit crunch?

The Fed reversed itself quickly, almost as though it were afraid of its own power. Monetary growth resumed, and so did the inflation. But for a brief period, the Fed held inflation in check with little help from the fiscal authorities. Because the full burden of price stabilization fell on the credit markets, it was very *destabilizing* in those markets. The shortage of lending capacity led many financial institutions to try to sell securities to get funds for their loan customers. This drove security prices down sharply. It became very difficult to sell newly issued securities. The holders of securities also saw the value of their assets melt away. This was particularly serious for financial intermediaries, whose assets fell relative to fixed liabilities. The financial community began to get jittery. There was talk of a ***financial panic***, which is a

speculative rush to turn securities into cash and to withdraw deposits from shaky financial institutions.

By the first week of September 1966, the situation was serious enough to make *The New York Times* devote front-page space to a poll of leading economists on the likelihood of a panic. Although none of them thought a panic was imminent, some thought it possible. And all were highly critical of the conduct of stabilization policy. They had two main complaints:

1. Monetary policy was much too expansionary in 1965, and then far too contractionary in 1966. The sudden turn-around was responsible for the near-panic conditions.
2. Placing the full burden of price stabilization on monetary policy while the budget remained highly expansionary unfairly distributed the costs. Some sectors of the economy suffered real hardship, while others were hardly touched. A tax increase would have been far more even-handed in its impact.

The benefit of hindsight has little to add to these judgments.

The tax surcharge of 1968

It would be wrong to think that the fiscal authorities failed to see the need to bring the growing inflation under control. The administration's economists were still Keynesians, and many members of Congress understood that the Keynesian lessons they had learned in 1963 and 1964 could be used against inflation as well as unemployment. The problem was to put together a fiscal program that would be stabilizing but also politically realistic. The program with the biggest following seemed to be simply to hope that the war would soon be over: Those were the days when the generals were forever seeing "the light at the end of the tunnel."

The administration first came forward with a major fiscal remedy in early 1967. It called for a personal and corporate income tax *surcharge.* A surcharge differs from a general tax revision in that it starts from the existing rate structure and just raises everyone's taxes by a uniform percentage. (Johnson's requested surcharge, in fact, exempted the lowest income taxpayers from any increase.) Such a program seemed likely to get through Congress more quickly than a general revision that would change the rate structure. It was to remain in effect "for 2 years or as long as the unusual special Vietnam costs continue(d)." Despite substantial congressional support, the surcharge was rejected in August. The President renewed his request the following year. In the face of both continuing warfare and inflation, the Congress passed the Revenue and Expenditures Control Act of 1968, which attached a 10 percent surcharge to corporate and personal taxes, and limited budget outlays. Since these spending limits specifically exempted Vietnam costs, interest, and Social Security, they were largely ineffectual.

Curiously, so was the surcharge. Understanding why should enrich your understanding of economic behavior. This is the main reason for looking closely at the surcharge.

The 1968 surcharge differed from the 1964 Revenue Act by being specifically temporary. It took effect in mid-1968 and was scheduled to expire in mid-1969. When the time came for its expiration, it was renewed for a year, but this could not be assumed when it was first enacted.

What seems to have happened is that consumers simply let their increased taxes come mainly out of income they would otherwise have saved. The share of disposable income saved, which was about 7½ percent in the year before the surcharge,

dropped to a little over 6 percent for the period when the surcharge was in full effect. This drop in saving almost exactly offset the rise in personal taxes. Apparently, the tax increase had little or no effect on consumption. This means it did little or nothing to slow the inflation.

There is good reason to suppose that a short-run tax change will be much less effective than a change that is expected to be permanent. A lot of personal saving is done by people who are building up funds for retirement. When taxes are permanently raised, people's incomes are reduced over the entire remaining years of their working lives. If they don't cut consumption, their retirement funds will be greatly impaired. But if the tax increase is temporary, its effects can be spread out over many years' consumption, without affecting any one year very much. Trying to control people's current consumption with a temporary tax change is like trying to push them around with a limp balloon. Most of the push just pops out somewhere else. The consumer response to the surcharge was very good evidence of this, but not very good stabilization policy.

Under the political circumstances of the time, it is hard to see how the fiscal authorities could have done much more. Fiscal policy is always governed by politics as well as economics. After all, those with the ultimate responsibility for fiscal policy *are* politicians. The country was in political turmoil over the war. A presidential election was due in November 1968, and it took some considerable political courage to impose a war tax at all, even if it was just temporary. The Democrats paid dearly for the war. Johnson decided not to run again. There were riots at the Democratic convention that August, and Nixon won the November election.

It is hard not to criticize the Fed, though. It was well insulated from the politics of the day. The economic situation of 1967–1968 called for continued monetary restraints. Interest rates would then have climbed high enough to crowd out private demand, keeping GNP from going way beyond its potential. Yet, after the credit crunch of 1966, the Fed veered way too far in the opposite direction, letting the M2 money supply increase by 10 percent in 1967 and by 9½ percent in 1968. Under the circumstances, this was not responsible public policy.

The Nixon-Ford years, 1969–1976

When President Nixon entered the White House in 1969, he inherited an "overheated economy," as the journalists like to call it. The unemployment rate was about 3½ percent. Consumer prices were rising by about 5 percent a year. Expenditures on the Vietnam War were at their peak, although new orders for military equipment were already declining. The major economic impact of the war was passing, but the economy had been operating beyond potential GNP for three years. The effects of excess demand were still working their way through the economy. Inflation had become a persistent and, to many, a serious problem. Nixon and his economic advisers were determined to bring this inflation to heel. They also shared a conservative belief that the federal budget should be balanced and limited in size.

The business cycle in the late 1960s and the 1970s

If you look at Figure 3, you will see what resembles a map of the Oregon coastal highway. The late 1960s were a continuation of the Vietnam period, with very low unemployment and mounting inflation. This was succeeded by a recession in 1970–1971 and a sharp drop in the infla-

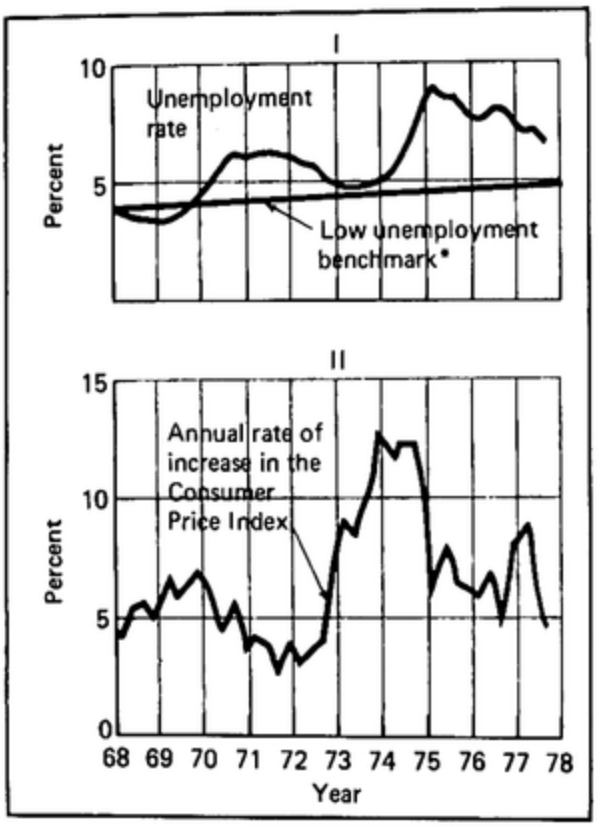


Figure 3 Unemployment and inflation under the Nixon and Ford administrations

The late 1960s and early 1970s displayed the familiar inverse relation between unemployment and inflation. But after the food and oil inflations of the mid-1970s, both unemployment and inflation were rampant.

*The unemployment rate defining potential GNP.

Source: Congressional Research Service, *Economic Stabilization Policies: The Historical Record*, 1962–1976.

tion rate. The economy recovered briefly from the recession, only to be battered by two massive "shocks"—the food inflation and the oil crisis. Prices climbed at a double-digit rate, real GNP fell, and unemployment rose to its highest level since the Great Depression. The economic crisis of 1973–1976 shook many people's faith in the continuing viability of American capitalism. It also shook their faith in Keynesian economics and in the economics profession. If ever there was a time to give a Council chairman nightmares, it was the mid-1970s.

Fiscal and monetary policies during the first administration

In attempting to control the inflation it inherited, the Nixon administration at first stuck to conventional policies. There was a surplus in the 1969 budget. When Nixon took office, the tax surcharge was still in effect, and he recommended its renewal in 1969. Despite efforts by a Democratic Congress to expand the War on Poverty, federal expenditures were held down. Nixon sometimes just refused to spend funds appropriated by Congress. Monetary policy was coordinated with fiscal policy, and for a while, everything went according to the textbooks. Before the end of Nixon's first term, inflation was subsiding.

The elements of fiscal and monetary policies are shown in Figures 4 and 5. The high-employment budget, which moved toward balance in late 1968 because of the surcharge, moved into surplus in 1969 as the surcharge stayed on and defense expenditures leveled off. At the same time, monetary growth dropped to nothing, in a replay of the 1966 credit crunch. Since there is no way to tell whether fiscal or monetary restraint triggered the rising unemployment, monetarists and Keynesians can both claim credit for the recession of 1970–1971, if "credit" is the right word. In any case, the rate of inflation fell off sharply, whether because of the budget, the money supply, or both.

Both monetary and fiscal policies were reversed in late 1970 and were expansionary in 1971 and 1972. With a lag, the economy recovered, and GNP reached its potential by the beginning of 1973.

The rate of inflation had dropped sharply in late 1970 and 1971, as unemployment rose to 6 percent. High unemployment caused prices to rise less rapidly than people expected and feared. From what you know about persistent inflation, expectations, and the shifting Phillips curve, you may reasonably suppose that as

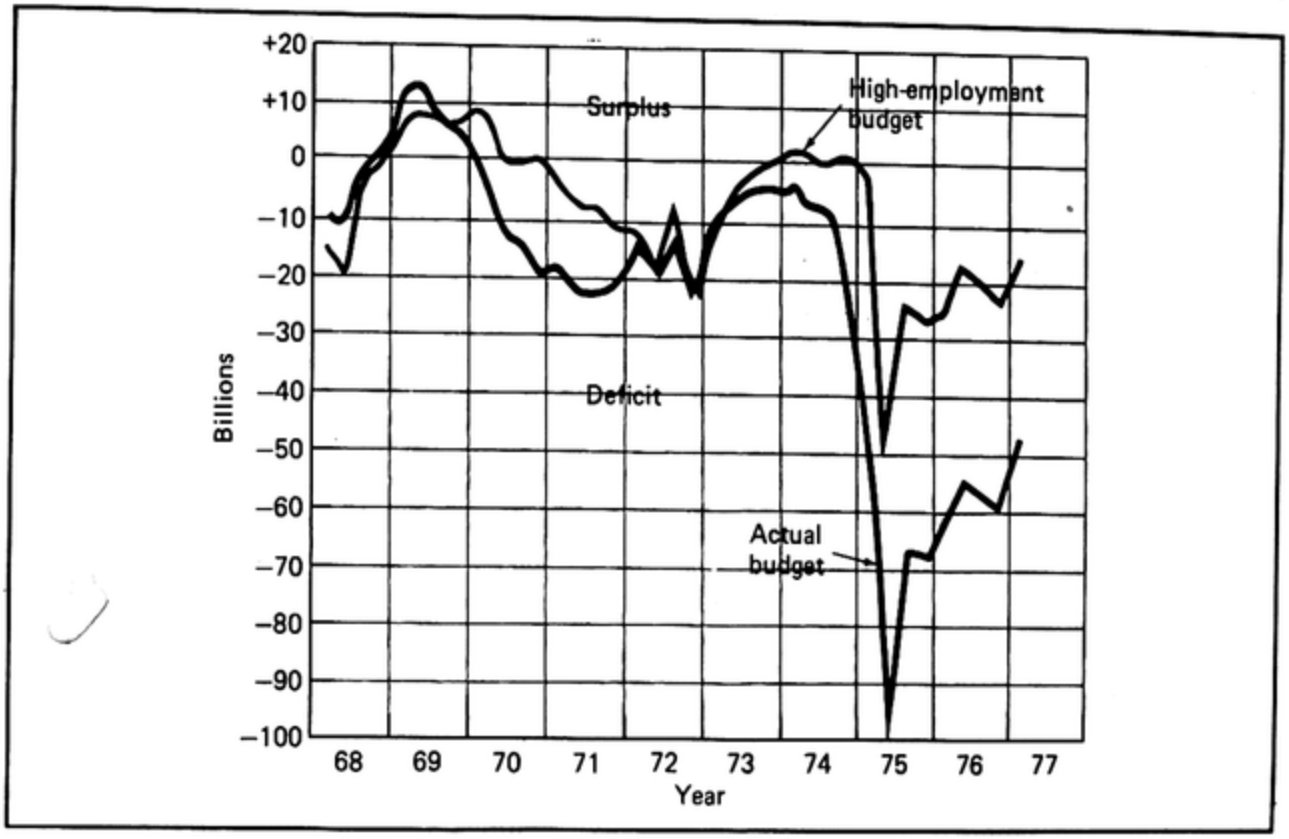


Figure 4 Actual and high employment federal budgets 1968–1977

The high-employment budget swung into surplus in 1969, reflecting the tax surcharge and a fall in military expenditures. The subsequent swings from 1970 through 1974 were aimed at stabilizing the cycle, but were comparatively limited. The enormous swing to deficit in 1975 was a large tax cut designed to fight the highest unemployment rates recorded since the Great Depression.

Source: Federal Reserve Bank of St. Louis, *Federal Budget Trends*, May 9, 1977.

the economy returned to potential, the rate of inflation would have stabilized at its late recession level, say 3 or 4 percent a year. Your guess is as good as ours. We'll never know what *would* have happened, because the drama was not allowed to proceed according to the books. In late 1971, the Nixon administration became impatient to deliver on its promises to end inflation. In a move that took most of the economy by surprise, it imposed ***direct controls*** on wages and prices, even though its policies seemed to be working well without them.

The period of direct controls

In the summer of 1970, Congress granted the President the authority to freeze wages and prices, despite Nixon's public opposition to such controls. The following March it extended this authority for another year. This time Nixon did not oppose the bill, but he showed no enthusiasm for it. Nor would you expect much enthusiasm from a generally conservative President. It therefore surprised most people when Nixon invoked the Economic Stabilization Act in August 1971, announcing a 90-day freeze on wages, prices, and rents. His move was part of a bundle of measures whose main goal was to cope with America's chronic balance-of-payments deficit. This deficit was partly the result of inflation, which continually raised American prices relative to those elsewhere in the world.

If you are ever President and plan to impose direct controls, it is a good idea to surprise everyone as Nixon did. Otherwise, they will try to get the jump on you. Tip-

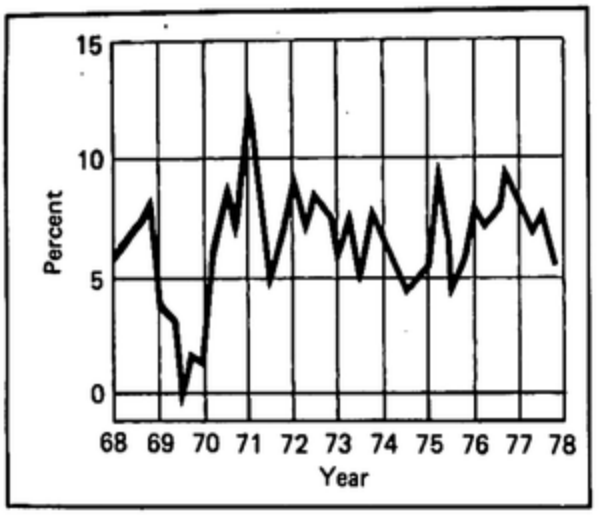


Figure 5 Annual rate of increase in the M2 money supply 1968–1977

Many of the quarter-to-quarter swings in the rate of monetary growth represent factors outside the Fed's control. But there was a pronounced period of monetary tightness just before the 1970 recession. Over most of the rest of the period, the money supply grew quite rapidly.

Source: *Federal Reserve Bulletin.*

ping your hand in advance only makes the inflation worse.

While the freeze was in effect, it stopped most price increases in their tracks, although it was not completely general in its coverage. From December 1970 to August 1971, the CPI grew at about a 4 percent annual rate and wholesale prices at about 5 percent. During the freeze, from August to November, the rate of increase in consumer prices dropped to 2 percent, and wholesale prices actually declined slightly. The administration was reluctant to fix prices for long, however. The freeze itself was invoked to buy time. It was followed by "Phase II," a set of complex, flexible, and largely voluntary guidelines. These do not seem to have been very effective. The Nixon Council itself referred to the first three months after the freeze was lifted as the period of the "bulge." Consumer prices increased at a 5 percent rate and wholesale prices at a 7 percent rate. You can see the freeze and the bulge in the bottom half of Figure 3. They appear as a dip at the end of 1971 and a peak at the beginning of 1972. Smoothing them out, it is hard to see much evidence of a marked decline in the inflation rate.

Phase II was succeeded by a series of other phases of varying degrees of strictness and ineffectiveness. The Council's *Reports* (which were required by the Stabilization Act to contain sections covering actions taken under the act) grew increasingly embarrassed about the whole episode. They tended to emphasize the costs of the program and to question its effectiveness. Finally, after the program had ended, the Council penned the following epitaph in its 1975 *Report* to President Ford:

> The final judgment on the effects of price and wage controls imposed under authority of the Economic Stabilization Act beginning in August 1971 and continuing for more than 32 months will be long debated and may never be resolved. The primary reason for an inconclusive judgment is that there is no way of accurately simulating the course of events which would have evolved in the absence of controls.
>
> However, the evidence of the controls period—including not only the behavior of the recorded rate of inflation but also materials shortages and other significant market events—does support a partial but important judgment about the experience with the controls system; regardless of the overall effect of the program, whatever contribution it may have made was probably concentrated in its first 16 months, when the economy was operating well below its potential. As various industrial sectors reached capacity operations in 1973 under the stimulus of a booming domestic and world economy, the controls system began to obstruct normal supplier-purchaser relationships, and in some cases the controls became quite unworkable. The sharply rising costs of basic materials, often reflecting world market influences and dollar devaluation, were largely uncontrolled; and when passed through to consumers they resulted in accelerating inflation. Thus, the net benefit of the controls system, however evaluated, had become extremely small by the beginning of 1974, and legal termination of controls only ratified the inevitable process of dismantling them in response to public and market pressures.

Milton Friedman once said that the imposition of direct price and wage controls was like smashing the thermostat because you don't like the temperature. Apparently, the Nixon-Ford administrations finally came to the same conclusion.

The second Republican administration

You already know a lot about the inflation of 1973–1974. By Inauguration Day in 1973, the big food inflation was already under way. The restrictive monetary and fiscal policy that seemed to work well in earlier years was ill suited to deal with the food and oil problems. All that is left is to tell how it broke down.

Soon after the inflation started in early 1973, fiscal policy became much more restrictive (see Figure 4), and the rate of monetary growth tapered off (see Figure 5). This, again, was the standard policy prescription, though the dosage was not very large.

There followed another recession, this time a very bad one. The unemployment rate reached nearly 9 percent in early 1975 and averaged 8½ percent for the year as a whole. It is hard to blame the recession entirely on the anti-inflation policy. The oil crisis in particular greatly shook the country's confidence, and some reaction would have followed, even if fiscal and monetary policies had not been directed at restraining demand.

What followed, of course, was a dilemma created by the very high unemployment in 1975 combined with the rippling out of inflation from the big splash in 1974. No stabilization policy looked good. What would you have done? President Ford (by this time Nixon had resigned over the Watergate affair) and Congress chose a tax cut of about $22 billion. Ford was also forced by Congress and the Supreme Court to release in the second quarter of 1975 appropriated funds that his predecessor had refused to spend in previous years. The high-employment budget plummeted into deficit and then rebounded in the third quarter to what was still a very expansionary position (see Figure 4). Unemployment declined, but remained a stubborn problem through 1976. So did inflation. The Republicans lost the 1976 election.

If there is any lesson from the 1974–1976 period, it is that fiscal and monetary policies cannot handle some situations. The big inflation of the mid-1970s was not simply a demand inflation. Shortages of food and oil caused price increases from the supply side, not from exceptional demand. Fiscal and monetary policies are most successful at ***demand management.*** They are quite general in their effects. When inflation originates in major *supply shocks* (like the food shortage and the oil cartel), general demand restriction is likely just to cause general recession without striking at the source of the inflation. Nor can it prevent the inflation from fanning out to other sectors of the economy. Cost increases will work their way through the input-output structure despite demand weakness. If the general price level is to be kept stable in the face of supply shocks, some prices must be *forced down.* The amount of unemployment necessary to do this when food and fuel prices are rising is likely to be very large. Since the cost of living is forced up by the supply shocks, there is considerable pressure for money wage increases, and it takes a lot of unemployment to keep them within bounds low enough for costs and prices in some sectors of the economy to fall.

The Carter administration

During the 1960s and 1970s, every incoming president had to face an immediate economic crisis of some severity. Jimmy Carter was no exception. In January 1977, the unemployment rate was about 7½ per-

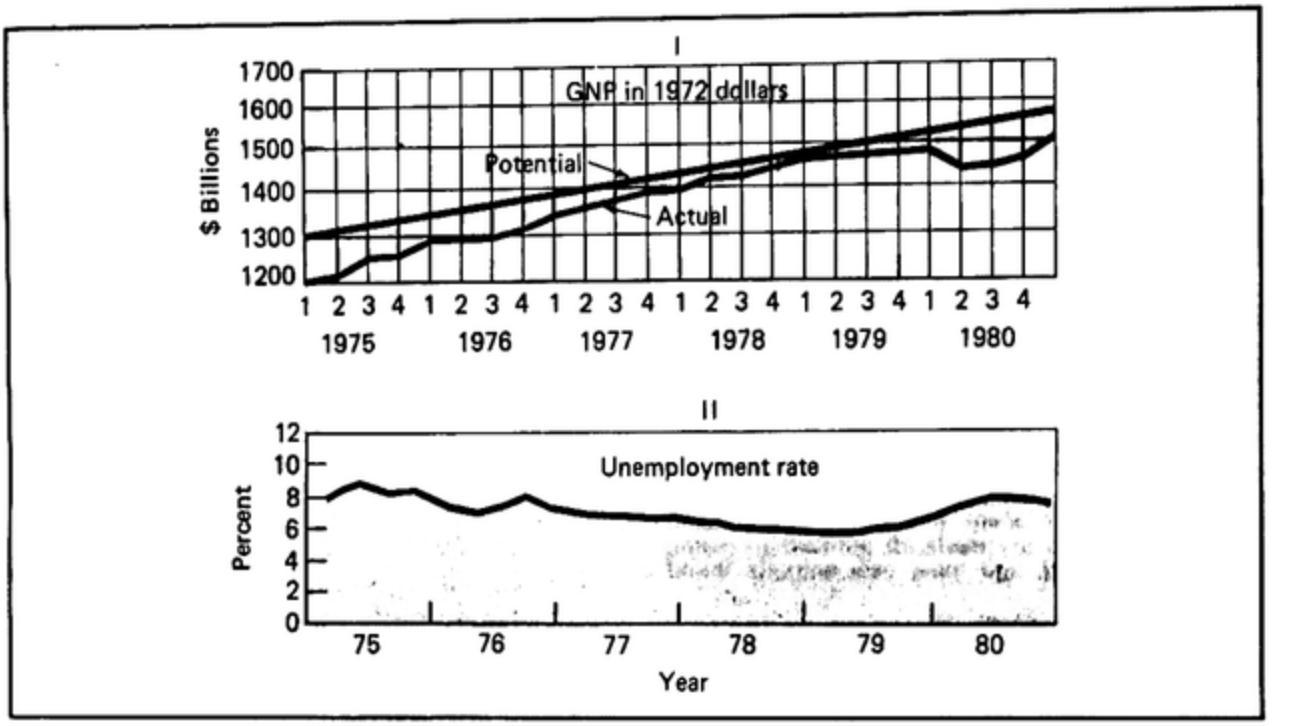


Figure 6 GNP and unemployment 1975–1980

During the Carter administration, the economy recovered slowly from the 1975–1976 recession. But the unemployment rate never dropped below 5½ percent, and there was another sharp recession during 1980.

Source: *Economic Report of the President.*

cent, and consumer prices were rising at about 7 percent a year. When he left office four years later, the unemployment rate was still about 7½ percent, but prices were rising at over 12 percent a year. Carter's years in office were marked by continued frustration of the government's efforts to contain inflation and reduce unemployment.

Prices and employment under the Carter administration

As you can see from Figure 6, in the first two and a half years of Carter's administration, unemployment was slowly reduced. By the summer of 1979, the unemployment rate had dipped to almost 5½ percent. It never got any lower, and by the end of that year, it was clear that another recession was developing. The recession of 1980 was not so bad as its predecessor—the unemployment rate topped out at 7½ percent. Unfortunately for Carter, however, the recession was badly timed. It reached its trough during the election campaign and contributed to his defeat.

The unemployment figures don't do full justice to what was happening in the labor market during the 1970s. If you look at Figure 7, you will see that from 1971 to 1980, employment grew much more rapidly than the population. In 1980, 58½ percent of the population was employed, up 2½ percentage points from a decade earlier. This growth in employment failed to cut into the unemployment rate because of a sharp increase in the fraction of the population actively seeking work. From 1948 through the end of the 1960s, this fraction—the *labor force participation rate*—fluctuated around 59 or 60 percent. A slow rise in female participation in the labor market offset a slow decline in male participation. During the 1970s, the uptrend

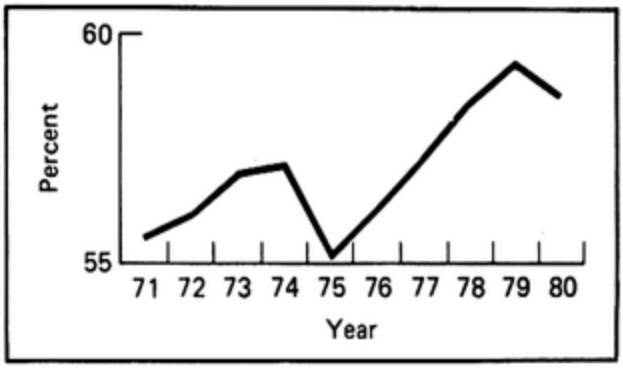


Figure 7 Employment as a percent of the population 1971–1980

Although unemployment was a serious problem during the 1970s and early 1980s, employment rose much more rapidly than the population during most of the decade. The share of the female population seeking work rose by nearly 10 percentage points from 1971 to 1980, and most new entrants found jobs.

Source: *Economic Report of the President.*

in female participation accelerated, and the total labor force grew very rapidly. Employment, in fact, grew quite fast throughout most of the decade, although it didn't keep pace with the labor force. This was true during both the Nixon-Ford and Carter years.

Carter fared somewhat worse than his predecessors in controlling inflation. The overall rate of increase in consumer prices jumped from 5 percent in 1976 to 7, 9, 13, and 12 percent in the succeeding years. As you can see from Figure 8, increases in food prices were a chronic problem, and energy price increases were staggering in 1979 and 1980. Prices of goods and services other than food and energy also grew quite rapidly toward the end of the decade.

All in all, the record of the American economy during the 1977–1980 period was a dismal one, despite the rapid growth in employment. The general public blamed the administration in office, and the popular press began increasingly to blame economists. The Keynesians, who were so cocky during the Lyndon Johnson years, began to dress in more humble garb.

Fiscal policy

By the standards of conventional fiscal policy, the Carter administration and Congress did a fairly good job of trying to contain inflation. As you can see from Figure 9, federal expenditures (on a high-employment basis) actually fell a little relative to potential GNP from 1977 to 1979. Meanwhile, federal taxes (again on a high-employment basis) rose quite sharply relative to potential GNP. The high-employment budget was in virtual balance by 1979. In 1980, however, sharp increases in transfer and interest payments opened up a substantial deficit again. Considering the amount of unemployment, fiscal policy still seemed quite restrictive.

In times of protracted inflation, though, conventional standards of policy may not apply the way they do when prices are more stable. In terms of current dollars, federal government expenditures went up by 56 percent from 1975 to 1980.

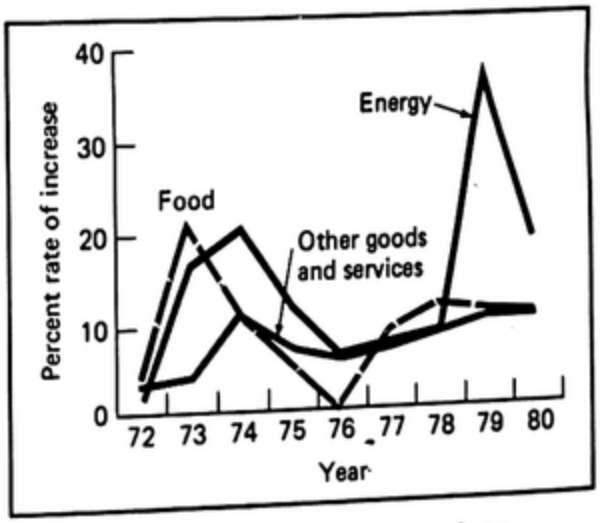


Figure 8 Increases in consumer prices 1972–1980

Following their sharp increases during 1973 and 1974, energy prices rose more slowly for several years. Then they jumped nearly 40 percent in 1979 and 20 percent in 1980. Food prices hardly increased at all in 1976, but rose about 10 percent a year in all four years of the Carter administration. The prices of goods and services other than energy and food were also increasing at 12 percent a year by 1980.

Source: *Economic Report of the President.*

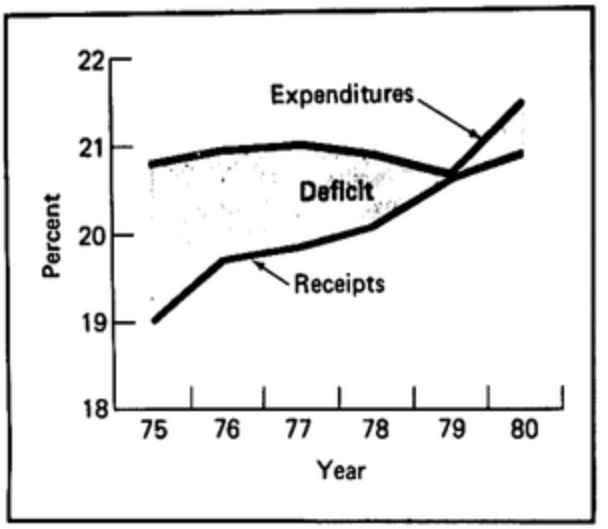


Figure 9 Federal high-employment receipts and expenditures as a percent of potential GNP 1975–1980

The federal budget grew increasingly tight following the recession of 1975–1976. In fact, the high-employment budget reached approximate balance in 1979, for the first time since 1974. In 1980, interest and transfer payments increased sharply, and tax refunds were unusually high because of the previous year's rate reduction. As a result, a deficit opened up in the high-employment budget.

Source: *Economic Report of the President.*

Most of this simply covered increases in the price level. But precisely because these expenditures kept pace with rising prices, they did little to stop the inflation, even though they were growing slowly in real terms. If the government continually builds a 7–10 percent inflation factor into its own expenditures, it can hardly be restraining the private sector from doing the same.

The tax side of the federal budget gets far better marks than the expenditure for controlling inflation. When prices and money incomes are rising rapidly, people move into higher tax brackets even though their real incomes are rising slowly or not at all. The particular treatment of depreciation expenses in corporate profits taxation also makes effective tax rates rise with inflation. If the fiscal authorities simply leave taxes alone, they will take a bigger and bigger share of real income as prices go up. It is this built-in property of the budget that was largely responsible for the rise in federal receipts relative to GNP that you see in Figure 9.

Monetary policy

The monetary authorities also moved to contain inflation during the Carter period. As you can see from Figure 10, there was a downtrend in the major monetary aggregates relative to GNP during most of the latter part of the 1970s. By this time, the Federal Reserve Board had become more impressed by monetarist teachings than it had been previously. In 1975, the Fed be-

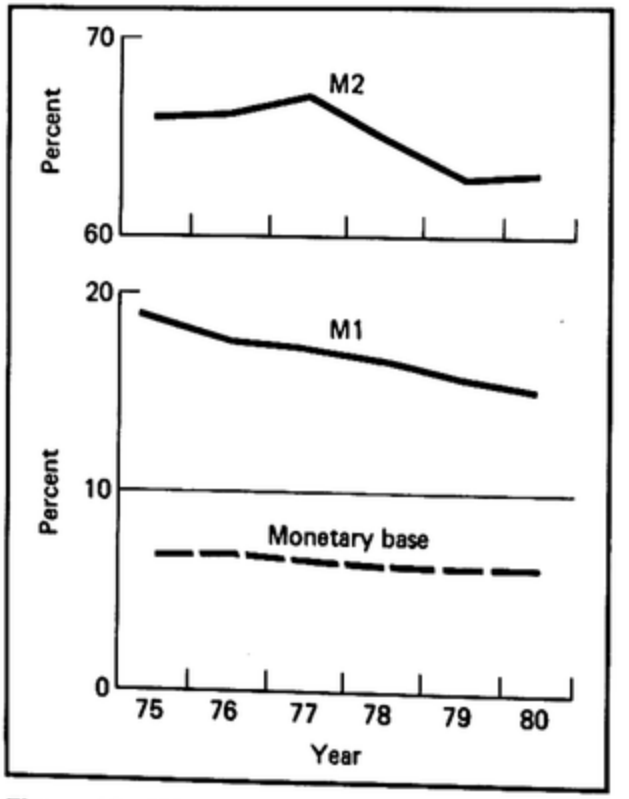


Figure 10 M2, M1, and the monetary base as a percent of current-dollar GNP 1975–1980

Except in 1977 (when an upturn in interest rates caused a large shift of funds into time deposits and a consequent increase in M2 relative to the monetary base), the Federal Reserve was successful in lowering the major monetary aggregates relative to GNP during the period between the 1975–1976 and 1980 recessions.

Source: *Economic Report of the President.*

gan announcing monetary targets publicly. As the decade wore on, it seemed to be trying to adhere closely to target rates of growth in the various measures of the money supply, and to be giving less attention to nominal rates of interest. Although monetarists complained that the Fed's monetary growth targets were too high, they must have felt some satisfaction at the recognition given to their theories.

Interest rates in the late 1970s soared to heights undreamed of in earlier decades, as you can see from Figure 11. If the Fed's anti-inflation policy can be faulted at all during this period, it is probably during 1977. It was slow to tighten money during the recovery from the 1975–1976 recession (note the bulge in M2 in Figure 10), so that interest rates didn't begin climbing until 1978. But when they did, they climbed fast, as Figure 11 shows. In fact, interest rates were permitted to climb higher in the recession year of 1980 than they had been in the peak year of 1979. A sharp drop occurred in midyear (see Figure 12), but rates soared again as soon as the recession passed its trough. They continued to soar into 1981.

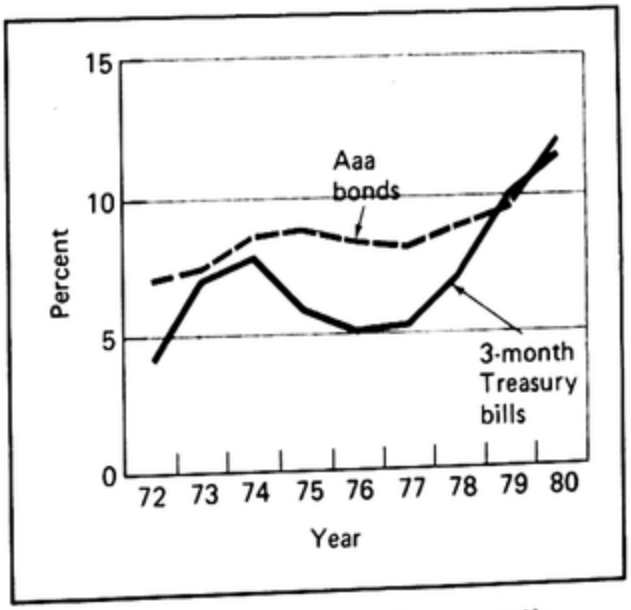


Figure 11 Interest rates on three-month Treasury bills and Aaa bonds 1972–1980

Interest rates followed their characteristic cyclical pattern during much of the 1972–1980 period, dipping in recession and climbing in recovery. However, they were higher in the recession year of 1980 than in the preceding 1979 cyclical peak. The monthly course of interest rates in 1979–1980 is shown in Figure 12.

Source: *Economic Report of the President.*

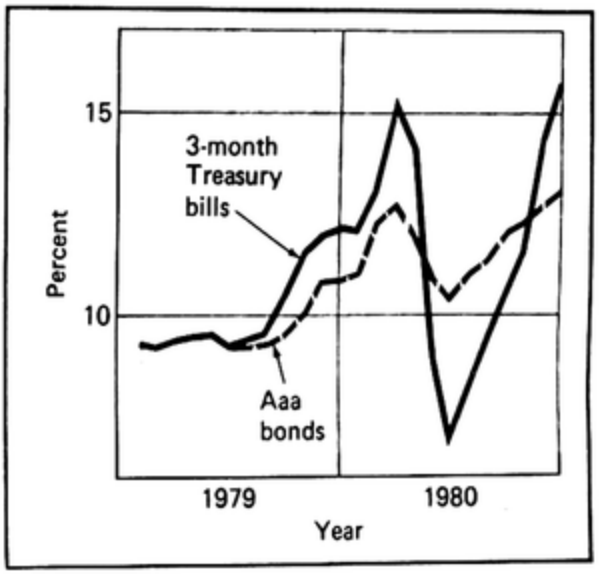


Figure 12 Interest rates during the 1980 recession

The dip in interest rates during the 1980 recession was sharp but extremely short-lived, so that for the year as a whole, rates were on average higher than they had been the year before.

Source: *Economic Report of the President.*

Was stabilization policy a failure during the Carter years?

Jimmy Carter was not a politically successful President, and he suffered for it at the polls in 1980. Economic conditions went from bad to worse during his administration. This is the opinion of almost everyone who has taken the trouble to look at the facts. But it does not follow that economic policy was conducted poorly during the Carter period. What, after all, were the alternatives? Within the range of what was feasible, there were only three:

1. Shoot for full employment, and let the price level go where it had to.

2. Tighten the fiscal and monetary screws so as to produce a "big bang," as it was then called—a recession big enough and long enough to stabilize prices.
3. Muddle through, trying to limit unemployment and bring down the rate of inflation.

The first alternative had almost no political support. There was support for the second, but no one seemed to know how big the bang had to be. Should we risk a Second Great Depression? Policymakers took the third course. If oil prices had been kept stable by OPEC and the world had been lucky enough to have a few good crop years, policy might have looked a lot better in retrospect. Under the circumstances, what would you have done?

The Reagan program

When Keynes' *General Theory* first appeared in the 1930s, someone wrote, "There is much that is new and good in this book, but what is good is not new and what is new is, unfortunately, not good." The same attitude greeted the economic program of President Reagan, who took office in January 1981. He was welcomed by a 7½ percent unemployment rate, and a level of consumer prices 12 percent higher than that of a year earlier. The Reagan administration responded with a policy program that combined orthodox monetarism with unorthodox views on budget policy. These latter views were called ***supply-side economics.***

Supply-side economics

The Keynesian tradition in stabilization policy emphasizes *demand management.* Growth in the labor force and productivity are largely (although not exclusively) taken for granted. The stabilization problem is to keep planned demand at an appropriate level relative to potential GNP, rather than to control the growth of potential itself.

Reagan's economic policymakers approached their new jobs with a different point of view. According to their analysis, the economic crisis of the 1970s stemmed in large part from systematic mismanagement of the supply side of the economy. They diagnosed the problem as slow growth in potential GNP. Although they freely acknowledged that production disruptions caused by high energy prices had contributed to the productivity lag of the 1970s, they primarily blamed the federal government for the slowdown in growth. Without giving specific figures, they ascribed a lot of the growth problem to excessive federal intervention in the economy. If the private economy were only released from its government shackles, they argued, potential GNP would have grown more rapidly. Because of their monetarist beliefs, they concluded that the maintenance of a stable monetary environment would be all the demand management necessary for prosperity, if only the supply side of the economy were free to operate efficiently.

It is hard to pin down exactly what any administration's policy is in detail, since policy statements are made by many people with varying degrees of influence. Some speak with the authority of the President behind them; others speak out of turn. Moreover, policy changes from time to time, partly because the economy gives changing signals, and partly because Congress sets changing limits on executive policy. (Remember the Rube Goldberg device.) But three major themes kept recurring in the Reagan administration's policy programs—*less regulation, lower taxes,* and *less government spending.*

The main argument for reduced regulation is that it increases productivity in the private sector. The government enforces a bewildering maze of price controls and supports, subsidies, pollution controls, and health and safety regulations. These directly benefit some people, but many economists think they cause a net social loss. The people who gain from them can see their benefits quite directly. Those who pay are widely dispersed and largely unaware of what the regulations cost them. Because of the asymmetry of information about gains and losses, the economy tends to be overregulated, the Reagan administration argued. It took aim at a great variety of regulations, both those favored by environmentalists and those favored by business interests. In doing so, it was only following the doctrines that free market economists like Milton Friedman have been putting forth for years.

The case for tax reduction was also an appeal to traditional economic arguments. Lower tax rates were expected to increase both work effort and capital accumulation. Reagan persuaded Congress to cut both corporate and personal taxes in the Economic Recovery Tax Act of 1981. The administration also argued for major revisions in the Social Security and welfare laws, mainly directed at strengthening work incentives. In particular, it advocated eliminating the "earnings test," which reduced Social Security benefits by 50 cents on every dollar of earned income above $6,000 per year.

Reducing federal expenditures was another integral part of the Reagan program. To some extent, the cuts were aimed at specific programs that were unpopular with Reagan's conservative supporters—such as welfare, subsidies, student loans, and nonmilitary research. But to a considerable extent, the cuts were also intended to reduce federal spending in general, regardless of its specific purpose. This would have two effects that were integral to the supply side program. First, planned demand would be reduced at each level of GNP. Second, resources would be released for employment in the private sector. With weaker consumption and government demand, the full-employment interest rate would drop, and investment would be stimulated. In effect, the reduction in government spending would reduce the extent of crowding out. The released resources would shift toward the production of investment goods, and the rate of capital accumulation would go up. This would affect the supply side by raising the rate of productivity growth.

In many respects, the Reagan program seemed internally consistent, although it was hard to tell because the administration was reluctant to spell out its goals and means in precise quantitative terms. Its failure to do this forestalled the kind of searching, analytical criticism that might have made the program more effective. The heaviest criticism came from social liberals, who attacked the pattern of income redistribution built into the 1981 tax and transfer cuts, which lowered taxes at the high end of the distribution and transfers at the low end. The 23 percent of households with incomes under $10,000 lost an average of $240 from the redistribution, while the 1 percent with incomes over $80,000 gained an average of $15,000. Keynesians wondered if the economy could attain full employment with a reduced federal budget. And economists in general were skeptical of claims that the program would greatly increase the growth rate of supply. But at least the plan as a whole was coherent. It appealed to monetarists and other fiscal conservatives, who expected the program of monetary restraint being pursued by the Fed to complement the shrinking budget. Prices

would stabilize, interest rates would drop, and private demand would expand to fill the gap left by lower government demand. Some important members of the administration were impressed by the arguments of the rational expectations theorists. Since they thought they were prescribing the right medicine, they expected the patient to recover quickly.

The defense budget

President Reagan campaigned on a platform that called for a stronger defense establishment. His budget director, David Stockman, thought this could be carried out within the framework of the government's overall economic plan, of which Stockman was the major architect. However, the Defense Department argued for a substantial program of new weapons development, and apparently won over the President. According to the 1982 *Economic Report of the President*, the administration's long-term budgetary plans called for a 9 percent annual growth rate in real military spending between 1981 and 1987. If accomplished, this will raise military purchases from 5.6 percent of GNP to 7.8 percent over the same period, and from 25 percent to 37 percent of the total federal budget. The implied budget deficits over the 1982–1987 period would average about $90 billion, or 2.4 percent of GNP.

Many economists and members of Congress have difficulty reconciling the military buildup with the administration's economic program. So, seemingly, does Mr. Stockman. There is simply not enough flexibility in the nondefense budget to permit this kind of military buildup in a context of overall budgetary contraction. If the Reagan administration can carry through its plans, the defense budget seems likely in the 1980s to be a major source of instability, as it was in the Korean and Vietnam War periods.

Postscript on policymaking

A couple of decades ago, a long-time Washington economist (who later became chairman of the Council) remarked jokingly that the country should have two identical capital cities. One would be for the government. The other would be for everyone else to visit. At the time, he was elbowing his way through crowds of tourists, trying to get to a meeting on time.

People with the power to make major policy decisions must often wish that the rest of us were far, far away. Surgeons who daily wield the power of life and death have to distance themselves from their patients as human beings. Otherwise, they could not stand the strains of their work. Economic policymakers are not so different. Cutting out health programs to trim the budget will kill people who otherwise would live. Inflation kills people. Every winter old people die of exposure because their pensions won't cover the fuel bills any longer. Unemployment kills people. Being unable to find work is frustrating and deeply degrading. It breeds domestic violence and suicide. Though the deaths may occur far from Washington, the dead are known in the capital—as statistics if not by name. Policymakers don't always think about this at work. Like the surgeons, they need distance. Yet, it must haunt them at night. They know that the "trade-offs" they deal in have human counterparts somewhere out there. It is an awesome responsibility.

Summary

This chapter has analyzed some of the triumphs and pratfalls of stabilization policy during the 1960s and 1970s. Its coverage is therefore selective rather than comprehensive. The historical period being

looked at presented policymakers with a series of problems to which they responded with varying degrees of success. You need not attach particular importance to any one problem unless it interests you. But you should carry away a few general lessons:

1. When the political situation is reasonably tranquil and the current stabilization problem is not very complex, it is possible to treat that problem scientifically and solve it. This was illustrated by the 1964 tax cut solution to the problem of underemployment in the early 1960s.
2. When the political situation is in turmoil, as it was during the Vietnam War, stabilization policy is so hemmed in by politics that it cannot be exercised in an effective manner, even though it seems clear what the current stabilization problem requires.
3. History sometimes presents policymakers with a genuine dilemma, as it did during most of the 1970s and early 1980s. Neither the experts nor the political powers have a clear view of what is to be done, and the conduct of stabilization policy is rent by contradictions.

Key concepts

Credit crunch
Tax surcharge
Financial panic
Direct controls
Demand management
Supply-side economics

Questions for review

1. a. Explain how a large high-unemployment, government surplus, coupled with weak private-sector investment, can prevent the economy from reaching full employment.
   b. What remedy for this situation did the Council of Economic Advisers propose in 1963? How was it expected to work?
2. The 1964 tax cut significantly affected consumer spending. The 1968 tax surcharge did not. Why was the latter so ineffectual?
3. Employment grew more rapidly than the population did from 1971 to 1980, yet unemployment remained high. Explain this seeming paradox.
4. a. Explain the difference in emphasis of *demand management* and *supply-side economics.*
   b. What are some supply-side arguments in favor of reducing government regulation, taxes, and federal expenditures? What are the counter arguments against such reductions?

33

American Economic Growth

As you read and study this chapter, you will learn:

- how important immigration was to the development of the American economy in the 19th and early 20th centuries
- why internal migration was also important
- what other demographic trends contributed to growth
- how modern agriculture and manufacturing industries developed in the 19th and 20th centuries
- why a national transportation network was crucial to overall development, and what role government played in its construction

Have you ever been puzzled by a medieval painting of the Virgin and Christ Child? There is always something vaguely wrong with it. The problem goes far beyond mere clumsiness in drawing the human anatomy: The bodily proportions of the child are entirely different from those of real children. It looks as though the artist deliberately tried to make him look like a miniature adult, scaled down on a one-to-five ratio. Fairly casual inspection will tell you that this is not how children are. The most immediate difference between an infant and a scaled-down adult is in the relative size of the head, but there are many others of no great subtlety. As children develop from infancy to adulthood, they go through many changes in muscular and skeletal proportions besides the obvious changes associated with sexual maturation.

Long-term economic change is a lot like human growth and development in this respect. Historical data on economic growth show evolution in economic proportions as well as changes in overall size. For example, agricultural employment in the United

States dropped from about 50 percent of total employment in 1870 to under 5 percent in 1970. This alone should convince you that growth is not just an expansion of scale. An examination of overall trends in GNP or any other broad aggregate conceals much of what is interesting about growth.

The study of economic growth is not the same as the study of economic history, which covers far wider territory. But growth itself is embedded in history and is a distinctly *historical* process. The upswing of a business cycle usually looks much like the preceding downswing, with the film run in reverse. Long-term growth, however, incorporates irreversible trends in science, technology, population, the division of labor in the workplace and in society, and much more. If the American economic system ever enters into a long decline and eventually collapses, you can be sure that it won't end with Columbus walking backward up the gangplank and waving good-bye to the Indians.

Most of the economics you have studied so far has taken a lot for granted. All of the major economic institutions and modes of behavior have been accepted as part of the given framework *within* which economic life is to be analyzed. Some examples are:

the structure and goals of households and firms;
the existence of markets;
the existence of nations and other political subdivisions;
the powers of courts, administrative agencies, and police;
the content of laws and legal precedents;
the content of traditions, beliefs, values, knowledge, common sense, and myth;
the state of technology and science.

These and many other aspects of society and human nature provide the backdrop for the analysis of economic developments over the short run. But all are subject to change over longer periods of time, and therefore cannot be taken for granted in studying economic growth.

Moreover, many people think that at least some of these aspects of society are profoundly affected by economic events. To the extent that this is true, it is not possible to study economic history only as a *reaction* to broader social change. It must also be studied as a *cause* of social change. The extreme version of this viewpoint is the *materialist conception of history*, which holds that economic developments set the tempo of history. Politics, law, culture, and even the structure of beliefs dance to the economic drum. Historical materialism is often misunderstood. It does not say that people make history simply to further their own immediate material ambitions. But it does say that how our economic lives are organized—how human society decides what, how, and for whom to produce—has a profound effect on how we conduct our politics and cultural life, and on what we believe in and fight for. In historical materialism, the method of production is assigned the pivotal role in history. But you don't have to believe that economics is pivotal to see how artificial it is to separate economic history from history in general.

The intimate connections among economic growth, economic history, and the history of politics, culture, and beliefs make it impossible to cover American economic growth in a single chapter. Most authors who devote even a whole book to the subject must sometimes despair at how hard it is not to be superficial. Thus, this chapter is very selective in its coverage. It addresses the process of industrialization in general, and in particular two major themes: first, the ***demographic changes***

from Colonial times to mid-20th century that gave this country an industrial labor force—the growth in population and the change in its composition and geographic distribution, and second, the process of ***capital accumulation and technical change*** that gave the United States its scientific farms, factories, and transportation system. Thus, it focuses on the growth in human and physical resources that have transformed the natural environment into a modern industrial country.

A word of caution about the statistics presented in this chapter: Remember the old adage, "There are lies, damned lies, and statistics," and always treat numbers with a healthy skepticism. This applies particularly to statistics describing what happened a long time ago. There are two reasons for this. First, since economic growth entails fundamental qualitative changes, the quantitative measures that are useful for describing one era are not necessarily right for another. Second, many of the events of earlier historical periods took place well before anyone systematically collected the raw data needed to measure them accurately.

Take GNP, for example. Nowadays, it is measured with great precision. At least, that is the view of the people at the Commerce Department who compute the statistics, and few economists seriously dispute their claim. Most debates over the GNP accounts center on comparatively minor matters dealing with the division of income into distributive shares, or else they are wholesale onslaughts on the GNP concept itself. Some attack it for leaving out home production, others for counting as output the expenses of coping with pollution, congestion, stress, and other by-products of modern life, and of waging war. But if you accept the Commerce Department's definition, you have to be fairly well pleased with the quality of the numbers in recent years.

However, nearly everyone who has studied the matter agrees that the GNP estimates for the past are much less good, and that the more distant the past, the worse the numbers are. Simon Kuznets was awarded the Nobel Prize for his work in reconstructing what national accounts would have looked like in the 19th century if anyone had bothered to collect the data. But you should use his figures as carefully as you would dynamite, the invention that made Alfred Nobel wealthy.

The two cautions about historical statistics apply with full force to long time series on GNP. The statistical raw material isn't there for measuring accurately the levels of market-oriented production a century ago. More importantly, however, the lines between production and other human activities were much less sharp in the 19th century. Today, most of us work from 7 to 3 or 9 to 5, and fit the rest of our lives around this block of time. In an earlier era, work was intermingled with the other activities of family life, and any split between the economy and the rest of society was very hard to make.

Of course, this sort of warning has about the same impact on economists as the surgeon general's warning has on cigarette smokers. They read it, sit back, and light up. GNP is such an important statistic for measuring economic growth that it is hard to ignore. Detailed estimates of constant-dollar GNP show a growth rate of a little over 4 percent from 1875 to 1914, a little under 3 percent from 1914 to 1949, and 3.6 percent from 1949 to 1975. The 3.6 percent figure is fairly accurate. The growth rate from 1914 to 1949 was doubtless lower than 3.6 percent, and the rate from 1875 to 1914 higher than the 1914–1949 rate. But who is to say that growth in the first of these three periods was really faster than in the third? You should think "fast, slow, fast," and not "4 percent, 3 percent, 3.6 percent."

Population

In 1670, there were about 112,000 colonists and slaves living in the North American colonies. By 1770, on the eve of the Revolution, the population of the colonies had increased to more than 2 million. By 1870, when the land area of the United States had reached roughly that of the 48 contiguous states, the country contained 40 million people. And by 1970, the 50 states had a population of 205 million. The rate of population growth over the 200 years from 1670 to 1870 was about 3 percent per year; over the following century, only about half as large.

Discussions of population and economic growth often begin with an outline of the *Malthusian population doctrine.* The *Essay on the Principle of Population,* first published in 1798, won for Thomas Malthus a fame that seems likely to last for centuries. In it, he argued that in a prosperous country, population growth would tend systematically to outstrip the growth of food production. The inevitable results of this tendency would be a growing scarcity of food, the spread of misery, and a check to population growth stemming from high death rates. Malthus' doctrine was systemized by his great contemporary, Ricardo, who combined it with the theory of diminishing returns to paint a bleak picture of the future. This picture was responsible for economics' coming to be known as "the dismal science."

Malthusianism is still a live doctrine, and it will probably become more prominent as the globe fills up with people. But the Malthusian tendencies were not among the major forces shaping American economic growth in this country's first centuries. Nearly empty land was abundant, and technological change continually staved off any trend toward diminishing returns.

Population and Immigration

The early colonists who arrived in Jamestown, Plymouth, and the other coastal settlements were a very small group relative to the native Americans already living in what we now know as the United States. But through a combination of military and political aggressiveness, economic expansion, and rapid population growth, the colonies grew to absorb the continent and to supplant the native population.

Early population figures are not very complete or reliable, but they seem to show a very rapid rate of population growth during the Colonial period up to the American Revolution. Most of this growth can be attributed either to ***immigration*** or to the fact that so many immigrants were young adults, who soon had many children.

The immigrants of Colonial times can be divided into three main classes, according to the circumstances under which they migrated: *free whites,* who were attracted by the opportunities of the New World or disenchanted with their opportunities in the Old and were sufficiently independent economically to afford their passage; *indentured whites,* who sold themselves into temporary servitude to get their passage from Europe paid by someone who needed their labor services; and *black slaves,* who came against their will, either directly from Africa or from the slave plantations of the Caribbean.

At the time of the Revolution, about 80 percent of the population was white. Approximately three-quarters of these people were of English, Welsh, Scottish, or Irish descent, although large numbers of Germans and Dutch lived in New York, New Jersey, and Pennsylvania. Of the 20 percent of the population that was black, most were slaves on the tobacco plantations of Maryland, Virginia, and the Carolinas.

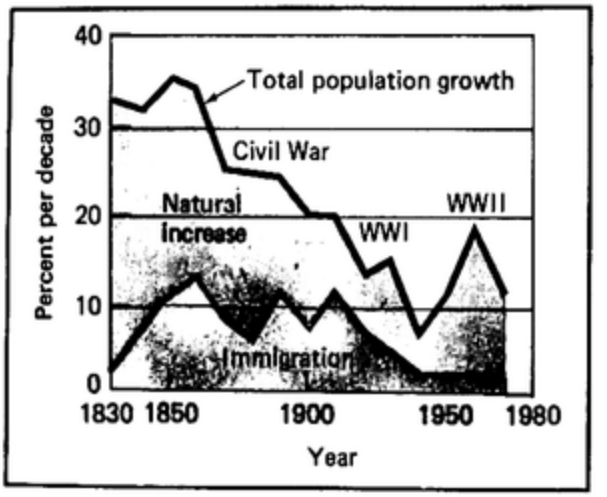


Figure 1 U.S. population growth 1830–1980
From the 1840s to 1920, immigration was a major force in U.S. population growth. In fact, in the first decade of the 20th century, growth due to immigration was more important than natural increase.
Source: *Historical Statistics of the United States.*

In 1820, the new United States government began to keep systematic data on immigration in addition to population census data. This makes it possible to see in detail how important immigration was to the country's growth in the 19th and early 20th centuries. Figure 1 shows that between 1830 and the Civil War, the population was growing between 30 and 35 percent per decade, doubling every 30 years. At first, much of this was due to ***natural increase,*** the excess of births over deaths. But as the century wore on, the rate of natural increase in population tapered off. Immigrants, however, began to pour in: 2 million Irish before the Civil War; 4 million Germans from 1850 to 1890; 6 million Central and East Europeans and 3 million Italians from 1900 to 1914. In the early 20th century, immigration once again outweighed natural increase, as it had during the Colonial period.

The national origins of European immigration from 1831 to 1929 are shown in detail in Figure 2. This immigration from Europe dwarfed the flow from other areas. Few Africans came after the slave trade was ended in 1808. Asian immigrants were important in some parts of the country, but not overall. It was not until well into the 20th century that Mexican immigration was sufficient to attract much notice, and it has always been local to the Southwest.

The immigrants, of course, went everywhere. But if you look at Figure 3, you

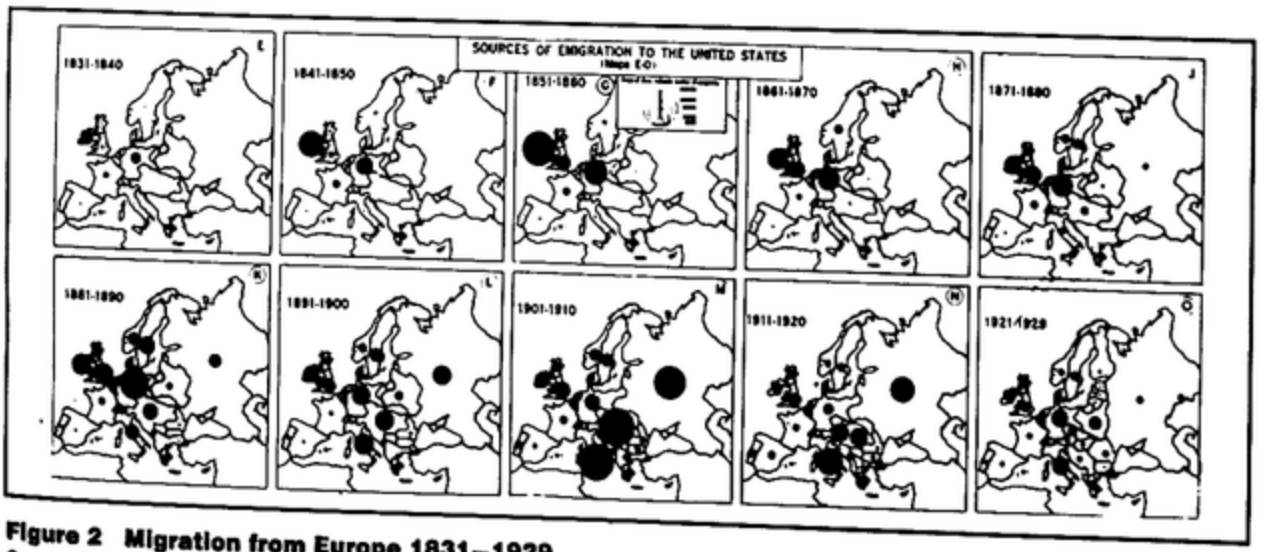


Figure 2 Migration from Europe 1831–1929
Source: Charles O. Paullin, *Atlas of the Historical Geography of the United States.* Published jointly by: Carnegie Institution of Washington and American Geographical Society of New York, 1932 (Carnegie Institution of Washington, Publication No. 401).

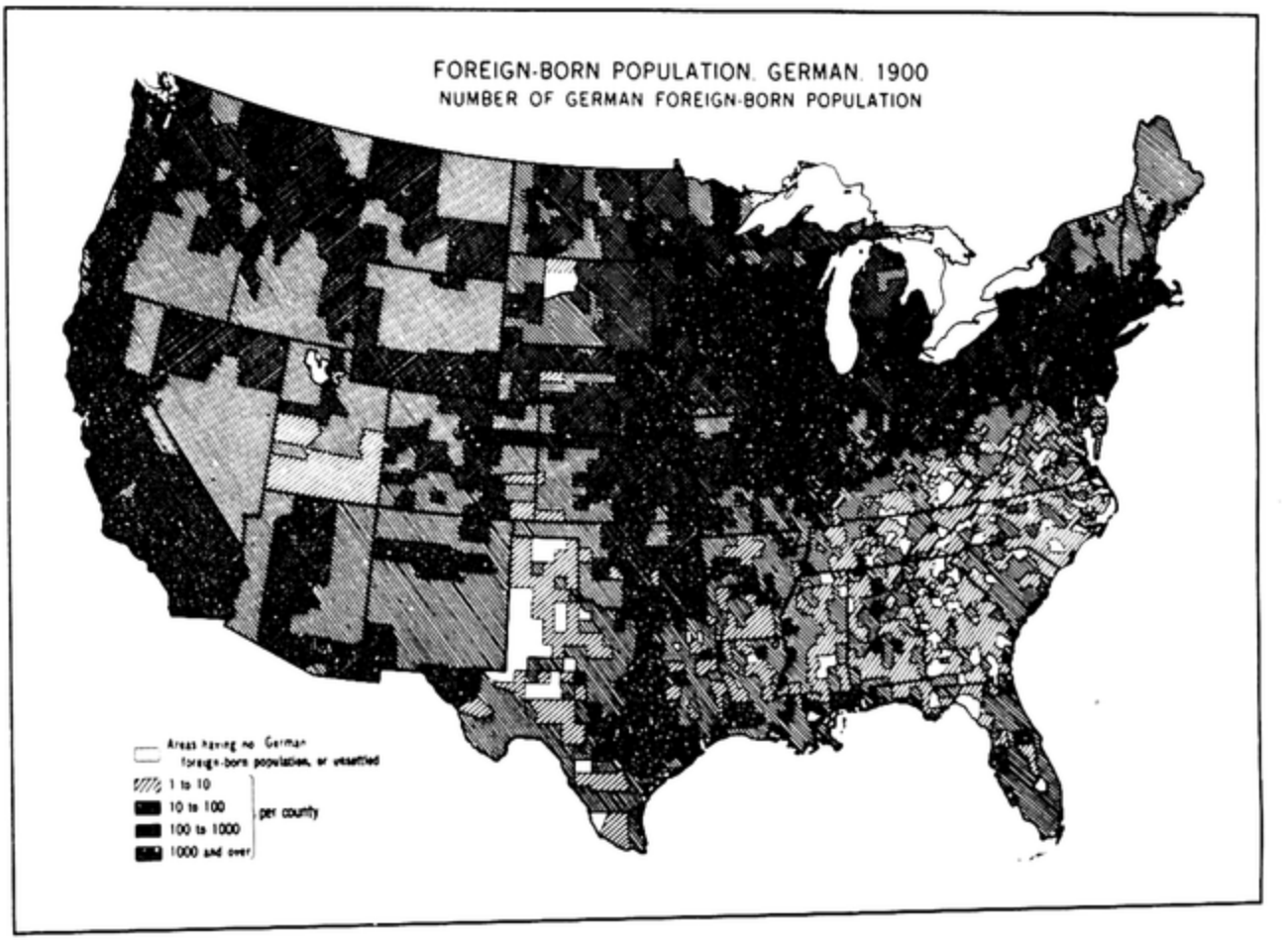


Figure 3 Foreign-born population, German, 1900

German immigrants settled mainly in the industrial Northeast and Midwest, and in the farming areas of the Missouri-Mississippi basin.

Source: Paullin, *Atlas of the Historical Geography of the United States*.

will see that the Germans, for example, settled mainly in the industrializing East and Midwest, and in the farming areas of the Missouri-Mississippi basin. Most of the other nationalities settled in a similar pattern. It largely reflected the economic forces that brought the immigrants here in the first place.

On the Fourth of July, orators tell us that immigrants came here seeking political and religious freedom. Ellis Island, the debarkation point for millions of 19th- and 20th-century immigrants from Europe, is part of a national monument that also includes Liberty Island. The inscription on its statue begins:

> Send me your tired, your poor, your huddled masses yearning to breathe free. . . .

The quest for freedom did draw many of the early Colonial settlers. It must also have drawn countless numbers who arrived much later. But on July 4, 1776, about one American in five was a slave, who either had come here in chains or was descended from those who had. And there were thousands of indentured servants who had sold themselves into servitude to get to the New World.

The great waves of white immigration can be accounted for by a combination of economic troubles in Europe and eco-

nomic opportunity in America—*push* and *pull*. A good example is the Irish immigration of the 1840s and 1850s. Potatoes in Ireland were struck by a blight that destroyed the principal occupation of the Irish peasants and produced widespread famine. At the same time, labor was scarce in the United States. The result was an enormous human resource transfer from Ireland to America. Two million people moved from where they were unproductive and poorly paid to where they were more productive and better paid. It is worth noting that thousands of Irish also emigrated to the slums of industrial England. Since England has historically been Ireland's oppressor, it seems clear that Irish sought jobs and food, not freedom.

Bringing immigrants to the New World was a big business. The owners of sailing vessels and (later) steamships actively recruited immigrants in Europe, advertising the opportunities available across the Atlantic. Like many hucksters, they promised more than they could deliver. Life in the slums of industrial cities and on the primitive farms of the Midwest was not pleasant. About a third of the immigrants eventually returned to Europe. But you will notice that the big money was to be made moving people from east to west across the Atlantic, not the other direction. As bad as immigrant life may have been, most of the immigrants seem to have been glad they came.

For rapid economic development, immigration is far superior to natural increase as a source of growth in the labor force. There are two related reasons. First, population growth by immigration lowers what the economic demographers call the ***dependency ratio.*** Every society supports many people who are neither employed nor actively productive in the economy of the household. Some of these are elderly, some sick, and some merely idle. But much of the dependent population consists of children. Like the lilies of the field, "They toil not, neither do they spin." But they can really pack away the groceries and wear shoes out in weeks. Productive workers support their own consumption with something left over, but the dependent population consumes without producing. The greater their consumption is, the less product there is for saving and capital accumulation. Thus, a high ratio of dependent population to working population (the dependency ratio) lowers investment per worker and per capita economic growth. A country whose population grows rapidly through natural increase always has a high ratio of children to adults and, therefore, a high dependency ratio. If the same high rate of population growth is achieved through immigration of adults, the dependency ratio is much lower, and the rate of economic growth can be correspondingly higher. Of the 1.3 million people who immigrated in 1907, about 85 percent were between the ages of 15 and 44, in their prime working years.

The second major reason for the superiority of immigrant labor is that some immigrants bring *human capital* with them—valuable skills beyond those that are normal among adult workers. In the 1930s and 1940s, for example, most of the leading classical musicians in this country were fugitives from European fascism, attracted by the security, freedom, and (yes) high incomes available to them in the United States. Now we train our own instrumentalists, singers, and even a few conductors, at great expense. But at one time, the best in the world flocked here by the score, at the height of their artistic powers.

The great immigrations benefited the immigrants themselves, their employers in this country, and the rate of economic growth of the United States as a whole.

Why, then, did Congress pass a series of laws in the 1920s that slowed immigration to a trickle?

The answer is not very complicated or surprising. It reflects the "pull up the ladder Jack, I'm on board" mentality of those already here. Americans of British descent organized as early as the 1840s to oppose Irish immigration. Their public statements were largely racist, but they must also have feared the loss of their jobs. Working people have always understood that continued immigration threatened their livelihood. In the late 19th century, employers who were faced with chronic labor trouble actually imported contract labor from Europe to replace their strike-prone workers. Working-class opposition to continued immigration came early. Yet, as long as employers favored continued immigration, Congress refused to halt it.

Two major events in European history finally spelled the end of free immigration. The first was World War I, which provoked strong anti-German feelings and general antagonism toward foreigners. The second was the 1917 Revolution in Russia, which sent a scare through the capitalist world. Immigrants from the European continent had long formed the nucleus of the labor and radical movements in the United States. After the Russian communists overthrew their government, Americans of property began to look for those who might overthrow the U.S. government. The anarchist under the bed, with his beard and bomb, always had an immigrant look. Employers threw in their lot with their employees. In 1924, immigration was restricted to about 150,000 a year.

It is one of the great ironies of American history that very soon this restriction became unnecessary. During the 1930s, the American economy collapsed so badly that masses of Europeans stopped wanting to come here, and the immigration quotas went undersubscribed.

Internal migration

On the eve of the Civil War, most of the U.S. population was located in the Northeast, the Southeast, and the Great Lakes states. At the beginning of the California Gold Rush in 1849, only about 3 percent of the population lived in what are now the Mountain and Pacific states. Over the next 100 years, however, the population of the Mountain and Pacific areas grew at about 4 percent a year, more than twice the rate for the rest of the country. Exceptional growth was not confined entirely to the Far West, though. Florida's population increased at $3\frac{1}{2}$ percent a year over the same century.

This rapid growth in Florida and the Far West was almost entirely the result of migration. In part it came from immigration. California, for example, gained about 2 million foreign-born residents between 1860 and 1960. But mostly, it reflected ***internal migration*** of native-born Americans.

The precise pattern of internal migration is hard to pinpoint, partly because it was complex, and partly because the data are poor. But the broad outlines of what happened are clear enough. Whites who came from Europe during the great immigrations of the 19th and early 20th centuries settled mainly in the Northeast and Midwest. As they were moving in, native-born whites were moving out. This pattern largely reflected competition in the labor market. Two developments were taking place at the same time. First, as a result of scarce land and increasing mechanization, agriculture in the East and Midwest was developing a surplus population. Second, cities in these same regions were industrializing. Without immigration from Europe, the surplus agrarian population might have migrated to these cities to form the new urban labor force, much as it had when Europe industrialized. But the European immigrants moved in, depressing wages and taking the new jobs. Thus,

the exodus from the farmlands moved west, where land was cheap, labor was productive, and farm wages were high. This westward migration was encouraged by the Homestead Act of 1862, which gave settlers the right to claim 160 acres of farmland for a nominal fee, provided they built a home on the land and improved it. Between 1873 and 1939, at least 1 million acres a year passed from public to private ownership in this way. The amount reached a peak of 10 million acres in 1913.

The second major internal migration was the exodus of blacks from the South. This movement was remarkably slow in the 50 years that followed the Civil War, but it accelerated during the 1920s and 1930s, World War II, and the prosperous decades after the war. A principal cause of large-scale migration was mechanization of southern agriculture, which drove many blacks off the land. From 1940 to 1970, blacks were attracted by the rapid growth of employment in northern industry. Like the migration from Europe, the black migration resulted from a combination of push and pull.

The net migration of blacks out of Alabama, Arkansas, Mississippi, and South Carolina exceeded natural increase, so that these states actually lost black population from 1940 to 1970. The main destination of black migration was the industrial North, particularly New York, New Jersey, Pennsylvania, Ohio, Illinois, and Michigan. In 1870, the black population of these states collectively was about 2 percent; in 1970, it was about 11 percent. By contrast, blacks were about 36 percent of the population of the South in 1870 and only 19 percent in 1970.

This complex pattern of international and internal migration greatly altered the relative populations of the major regions of the country. As you can see from Figure 4, in 1800, the Northeast and South between them just about split the population

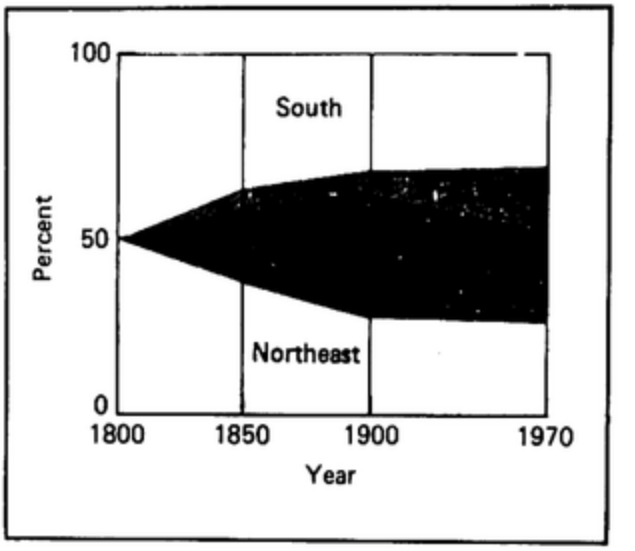


Figure 4 Changes in the share of the U.S. population living in major regions of the country 1800–1970

During the 19th century, international and internal migration greatly increased the relative importance of the north-central region. Beginning in the mid-19th century, the West began to grow rapidly. In the 20th century, it has increased in population relative to all three other major regions.

Source: *Historical Statistics of the United States.*

of the area that is now the United States. By 1850, the north-central region had a quarter of the population. The West was still relatively unimportant. In the second half of the 19th century, the West and the north-central regions both grew relative to the two regions that had dominated the population map in 1800. In the 20th century, the West has grown relative to all these other regions. This last trend continues today.

Urbanization

At the first U.S. census, in 1790, about 95 percent of the population lived in rural areas and very small towns. Beginning in the 1820s, and gathering momentum from the European immigration, the United States was gradually transformed from a rural to an urban nation. This trend is illustrated in Figure 5. By 1900, the Northeast was largely urban. By 1970, all four of

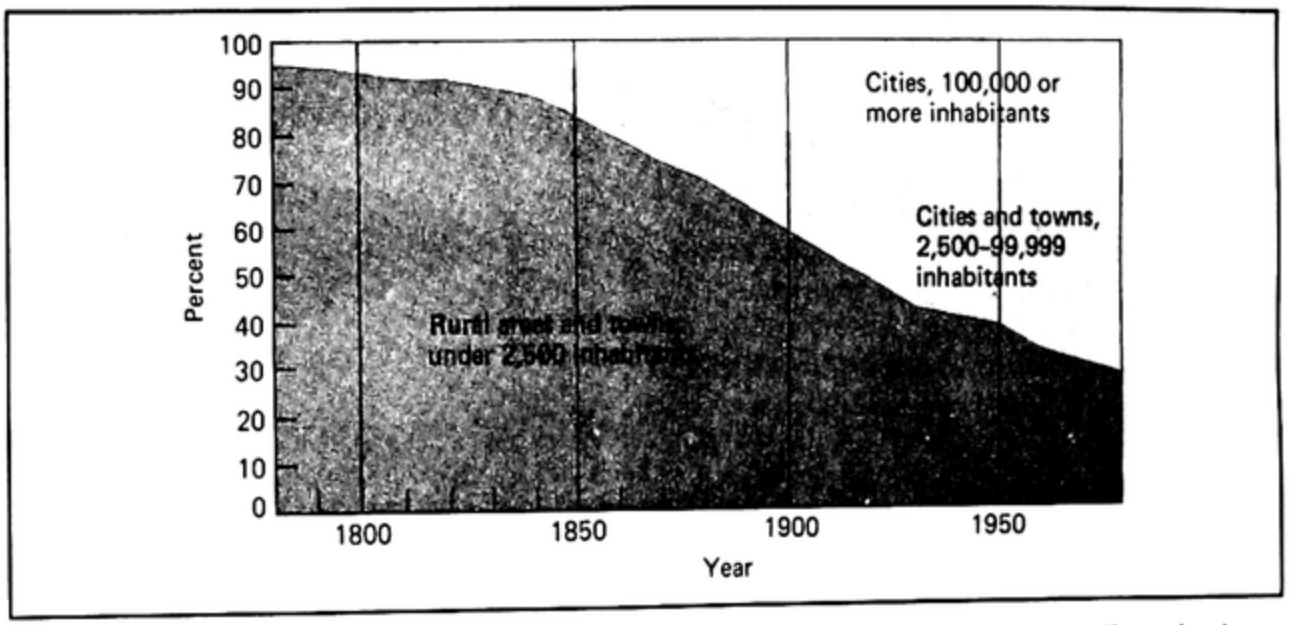


Figure 5 Distribution of the U.S. population among rural areas and small towns, medium-sized towns and cities, and large cities 1790–1970

The growth of the U.S. population has been accompanied by an increasing concentration in urban areas.

Source: *Historical Statistics of the United States.*

the major regions were more than half urban, although the South still had 35 percent of its population in rural areas. Curiously, the West, which is the most sparsely populated region of the country, is also the most urbanized. In 1970, 83 percent of its population lived in urban areas. The other 17 percent were thinly sprinkled over its vast forests and prairies.

At the time of the American Revolution, London had a population of nearly 1 million. It was not until the Census of 1820 that the United States could boast a single city (New York) with more than 100,000 inhabitants. In that year, the four great cities of the East Coast (Boston, New York, Philadelphia, and Baltimore) contained nearly half the country's urban population. The ***urbanization*** that took place in the 19th century was a flowering of new cities, not simply a further expansion of the old ones. By 1920, the four largest cities (New York and Philadelphia still, but now Chicago and Detroit) had only about 20 percent of the urban population. The growth in the number of large cities is shown in Table 1.

The development of new urban centers was part of the general pattern of economic growth. One major factor was the spread of agriculture into the Grain Belt of the Midwest, the Cotton Belt of Alabama, Mississippi, and Louisiana, and the cattle ranches of the West. International and internal commerce were no longer exclu-

Table 1 Growth in the numbers of large cities in the United States, 1790–1970

	Number of Cities with Population of:		
Year	1 million or more	500,000 –1 million	100,000 –500,000
1790	—	—	1
1820	—	—	5
1850	—	1	16
1880	1	3	32
1900	3	3	78
1940	5	9	130
1970	6	20	

Source: *Historical Statistics of the United States*

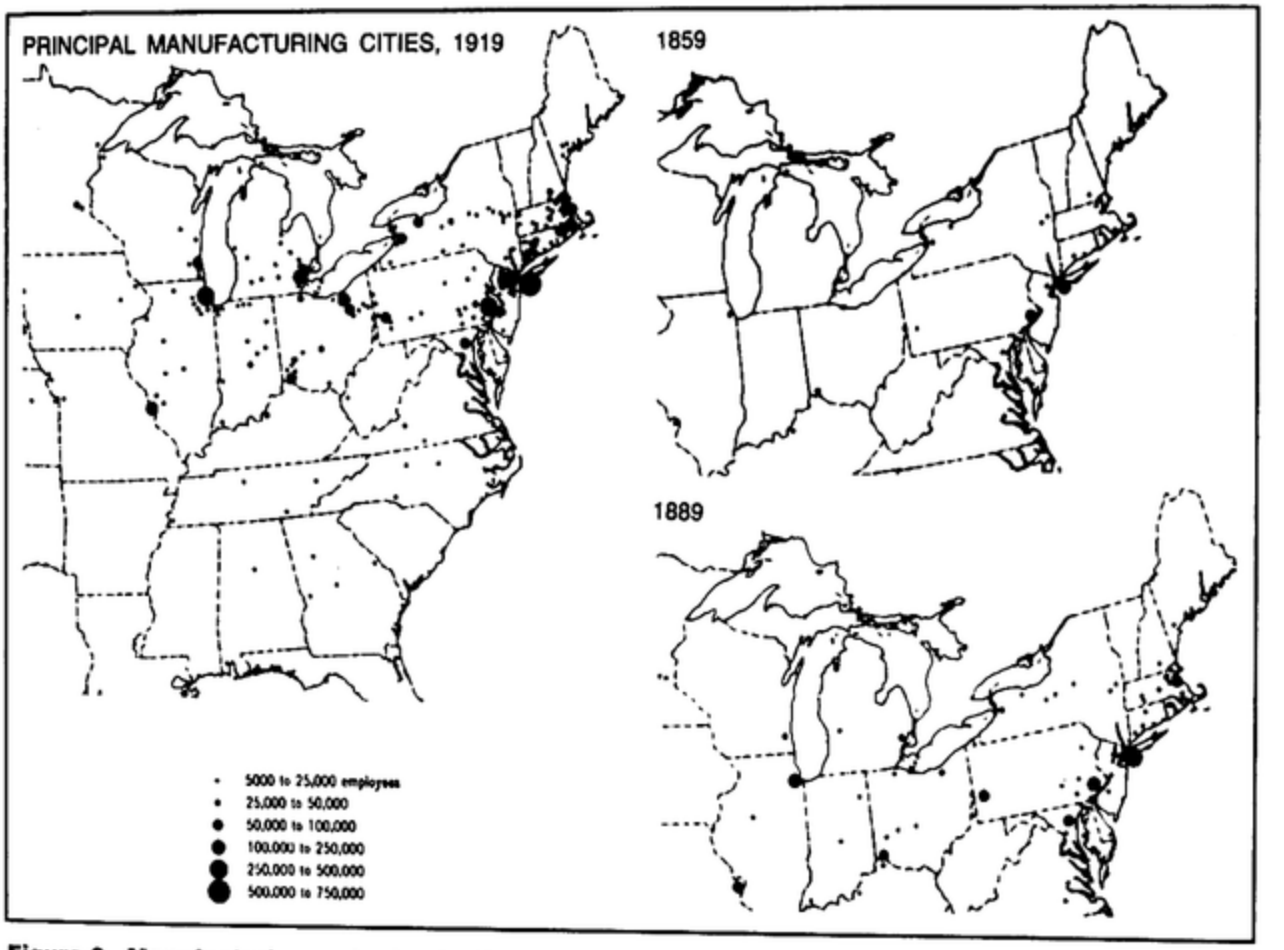


Figure 6 Manufacturing and urbanization
Much of the impetus to urban growth came from the spread of manufacturing across the eastern half of the country.
Source: Paullin, *Atlas of the Historical Geography of the United States.*

sively organized around the major East Coast cities. Pittsburgh, Cincinnati, Detroit, Kansas City, St. Louis, New Orleans, Chicago, Denver, San Antonio, and San Francisco became major centers of trade. A second major factor was the development of manufacturing. The textile and shoe industries grew up mainly in the settled areas of New England, which had water power and easy access by ocean transport to materials and markets. But the iron and steel industry, the meat-packing industry, and other industries whose access to materials and markets depended on inland roads, railroads, and waterways developed west of the Appalachian Mountains. Many of the major cities of the country—St. Louis, Chicago, Detroit, Cleveland, Cincinnati, Pittsburgh, and Buffalo—owe their industrialization to favorable locations on the Great Lakes or inland rivers. The growth of the principal manufacturing cities in the northeastern quarter of the country is shown in Figure 6.

Other demographic changes

Most Western philosophers of the past two centuries have proclaimed equality for all people. But industrialization is explicitly

choosy. It favors free labor of working age, in good health, willing to live in cities, work outside the home, and subject itself to the discipline of large-scale enterprises. It also requires some workers who are well enough educated to run these enterprises and develop their technology.

So far, this chapter has dealt with population growth, migration, and urbanization. But other demographic changes have also contributed to U.S. economic growth and industrialization. Most of them are illustrated in the various panels of Figure 7. You should not imagine that they are independent causes of economic development. Like population growth, migration, and urbanization, each is an effect as much as a cause. Demographic changes are part of the self-reproducing process of economic development.

Look first at the freedom of the labor force. Slavery reached its peak relative to free labor around 1800, when about 28 percent of the work force were slaves. With the ending of the slave trade in 1808 and the upsurge of European immigration, the slave population dwindled relative to the free population, so that by emancipation in 1863, it was only a little over 20 percent of the labor force. Emancipation did not immediately produce much labor mobility, however. The southern black labor force was too ill educated and poor to do much but work the land and perform domestic service, much as it had under slavery. But the free blacks were prepared to move by the time northern industry beckoned in the 20th century.

A second major factor in the development of workers suited to industrial growth was the change in the number of people who were in the labor force. This is a complex matter involving not only the composition of the immigrant population, but also changes in the age structure of the native-born population and in customs about work outside the home.

Consider first what it means to be "in the labor force." In Coionial times, when most work was family farming and handicraft, the separation between work and family life was less rigid than it is now. True, there was usually a division of labor in which men produced for the market and women for home consumption, but the two kinds of activities were intertwined and carried out in much the same place. As capitalism and its factory system developed, they took over many activities that had previously gone on in the home and centered them in capitalist enterprises. Thus, there developed two separate spheres, one of work for pay and the other of work in the home. The first was largely the province of men and unmarried women, the second of married women.

As you know, most home production is not customarily counted in GNP, and unpaid family workers are not counted as part of the labor force unless they participate directly in a family business. No matter how hard women work in the home, they are not part of the "work force" unless they also work outside.

One of the most important demographic trends of the past century or so has been the increase in the proportion of married women and women with children who work outside the home. In 1870, only 15 percent of the work force was female. Most of them were young and as yet unmarried. But by the end of the 1970s, a bit more than half of all women over the age of 16 were in the labor force, compared with about 75 percent of men. By this time, the labor force was more than 40 percent female.

A number of factors accounted for the rising participation of mature women in the labor force. One was the drop in the birth rate, a long-term trend that has reduced the amount of her life a woman spends either pregnant or nurturing infants. Second was the spread of free, com-

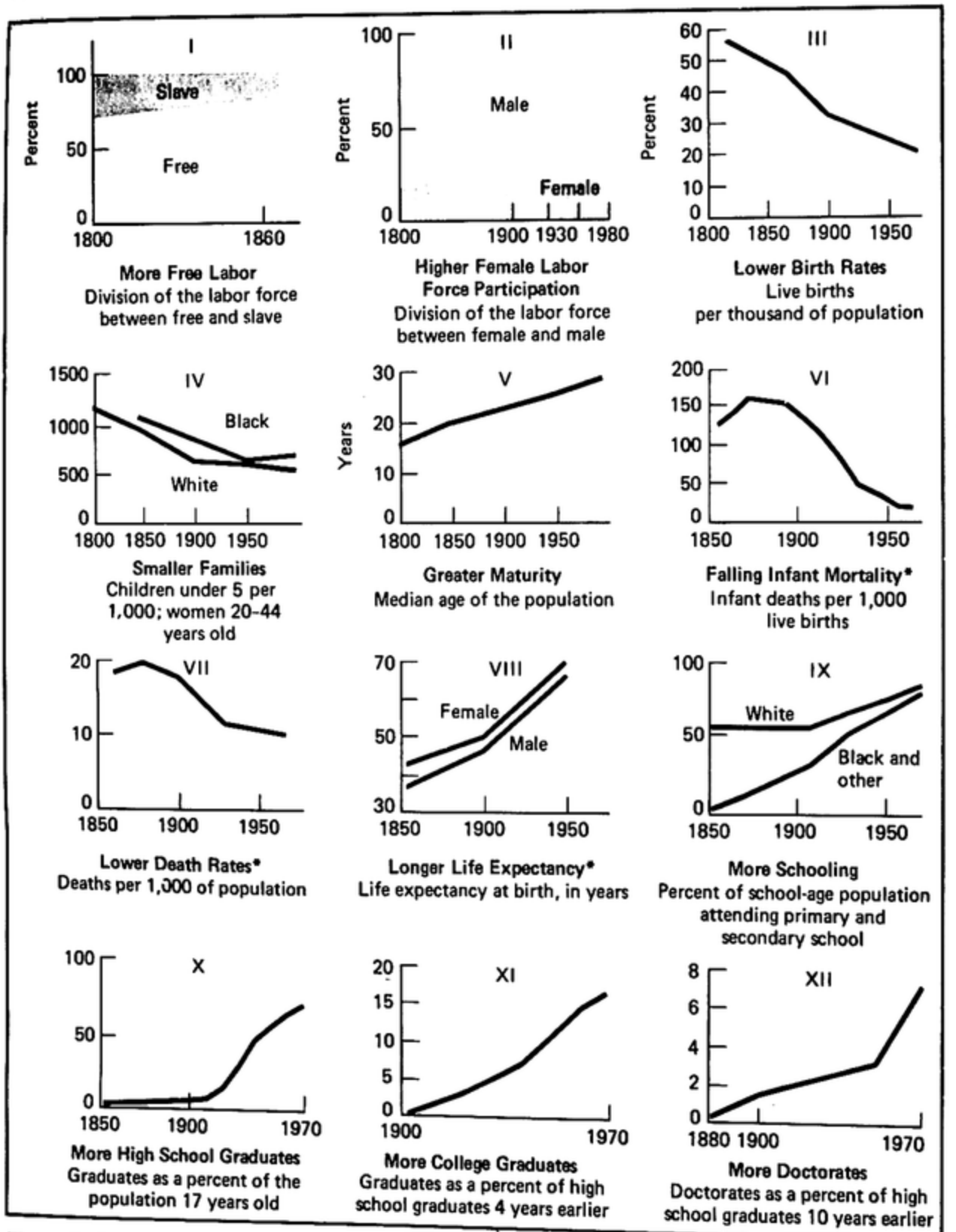


Figure 7 Demographic changes influencing the rate of economic growth

*Panels marked with an asterisk are based on data for the state of Massachusetts, which gathered quite complete demographic data before this was the custom in the rest of the country. It is hard to know how closely the Massachusetts data approximate the characteristics of the country as a whole. Yet, where overlapping data are available, Massachusetts demographic characteristics are roughly comparable to those of whites in the rest of the country.

Source: *Historical Statistics of the United States.*

pulsory public education, which made it easier for mothers to be absent from home for extended periods. Third was the development of household appliances and prepared foods, which made housework more flexible in its timing, and perhaps less burdensome. Fourth—partly related to the first three—was the rising divorce rate, which has made more and more adult women responsible for supporting themselves and their children. Finally, the exceptional growth in participation of married women during the 1970s doubtless reflected the fall in their husbands' real wages.

This rising participation rate for women was one element in a general downtrend in the dependency ratio. Another element was the uptrend in the average age of the population. In 1800, half the population was under 15 years old. Although many of these young people were useful hands on their families' farms, they would not have been nearly so useful as members of an urban labor force. But a falling birth rate, increased life expectancy, and adult immigration combined to raise the median age from 15 years in 1800 to 28 years in 1970. This contributed to a drop in the dependency ratio and enabled young people to spend more time in school without making major inroads on overall labor force participation.

The falling dependency ratio during much of the 19th century made a double contribution to growth. First, as the dependency ratio fell, the labor force rose more rapidly than the population as a whole. This contributed directly to growth in per capita output. Second, a *falling* ratio led eventually to a *low* ratio. As you know, a low dependency ratio contributes indirectly to growth by freeing resources for capital accumulation.

Another major demographic trend contributing to growth was the improvement in health, especially with the application of science to public health in the late 19th century and to medicine in the 20th century. The major contributor was the construction of urban water purification and sanitary sewer systems around the turn of the century. Immunization was developed against several major diseases. Antibiotics rendered innocuous most kinds of infection that had previously been dangerous. Improved surgical techniques made it possible to repair injuries that had once been crippling or fatal. Obviously, a healthier population is a more productive population. Almost as obviously, a population in which nearly everyone born survives to have a lengthy adult life span has (up to a point) a lower dependency ratio than one in which most people die young. The 20th-century United States has benefited from a falling death rate (especially infant mortality) and a rising life expectancy.

Another important 20th-century trend has been the growth in education. (For blacks, this trend began soon after the Civil War.) At the turn of the century, less than 10 percent of the population finished high school. The fraction is now approaching 80 percent. College attendance has grown relative to high school attendance, and the number of doctorates has grown relative to the number of bachelors degrees.

This trend toward greater educational attainment has contributed to economic development in three ways. First, most modern jobs require basic cognitive skills like speech, reading, understanding, writing, and arithmetic, besides elementary motor skills like pushing, pulling, lifting, and carrying. A better-educated work force is better equipped to deal with the cognitive component of ordinary work. Second, some kinds of jobs—such as scientific work, engineering, and higher-level management—require highly developed cognitive skills. An economy that produces

many highly skilled people can be better managed and more technically progressive than one that does not.

The third contribution of education is noncognitive. If you think back to your own experience in primary and secondary school, you will probably recall that a lot of effort was devoted to teaching you to show up on time, sit still, and obey without asking why. The ability to behave in this way is not something we have when we enter kindergarten. To a great extent, we learn it in school. And it is the exact mode of behavior required of most people who work in the 20th century factory or office. A society in which most people finish high school is one whose work force is well adapted to the discipline of modern enterprise.

Capital accumulation and technical change

The theory of growth

The overall relationship between demographic change and economic growth is complex. When people's health and education improve, any resulting increase in their productivity adds to per capita output. That much is obvious. Internal migration from labor-surplus areas to areas with a shortage of labor also adds to growth. But sheer increase in the size of the population may promote either growth or misery. If the country is so underpopulated that its labor force is not very specialized, population growth will add to productivity by extending the division of labor. Adam Smith recognized this explicitly. But his followers, Malthus and Ricardo, also recognized that if population increases relative to land area, the result will be diminishing marginal productivity of agricultural labor and eventual impoverishment. Sustained economic growth is impossible with land and labor alone. The inputs supplied by the limited natural environment must be supplemented by tools, machinery, and sources of power. Such capital inputs can sustain growth in per capita output even though the labor force is encroaching on a limited supply of land.

But there are also diminishing marginal returns to the use of capital. Capital of a given kind becomes redundant when its ratio to labor and land becomes too high. The piling up of more and more capital cannot compensate for a limited natural environment, and growth is destined to grind slowly to a halt unless people can continually become more clever in how they organize their labor and capital inputs. Continual technical change offers a way out of the trap of diminishing returns.

Economists often discuss the theory of growth with the help of an aggregate production function, which relates output to inputs of land, labor, and capital and to the state of technology. One way of representing the aggregate production function is illustrated in Figure 8. The curve shows that increases in capital per worker raise output per worker. The declining slope reflects diminishing marginal returns from an increase in the capital-labor ratio. When capital and labor both grow relative to the natural environment, the curve shifts down because of diminishing returns from working the land more intensively. Technical change shifts the curve upward, nullifying the diminishing returns from increases in capital and labor.

The lesson of the theory of growth is that a country that sustains economic growth does so by improving its productive techniques and increasing its capital-labor ratio. Even though this lesson is almost self-evident, it is very valuable because it tells us to look for the sources of growth in capital accumulation and technical improvement, and for the barriers to growth in more intensive land use.

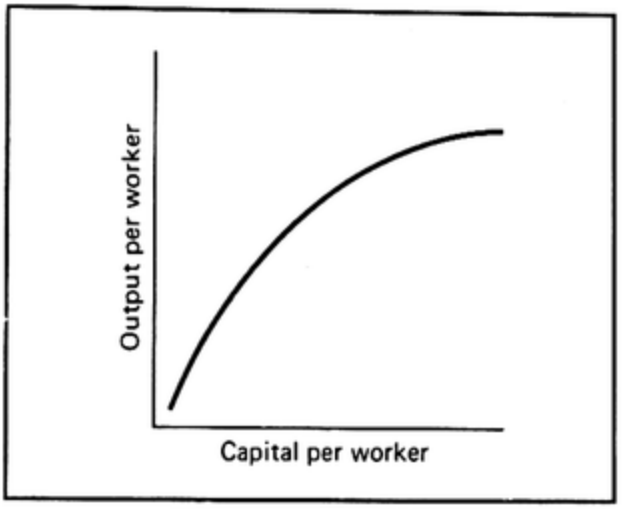


Figure 8 The aggregate production function

The declining slope of the aggregate production reflects diminishing returns from increases in the capital-labor ratio. Diminishing returns from the application of more capital and labor to a fixed natural environment are represented by a downward shift in the production function. Technical change is represented by an upward shift.

If output, capital, land, and labor were really four homogeneous substances, like the physical elements, it would be easy to parcel out growth in homogeneous output into what is due to changes in the quantities of the homogeneous inputs and what is due to improvements in their quality. As things are, however, output, capital, labor, and land are categories, not things. The particular occupants of those categories change over time. Most of what we produce now was unheard of 200 years ago, and much of what was produced then can only be found in a museum. The same goes for the capital goods we use and for the labor techniques we apply. The tools, techniques, and products that you can see in the living museum of Greenfield Village, Michigan, are almost totally different from those you would encounter in Ford Motor Company's River Rouge Plant a few miles away. Over long periods, quantity and quality are inseparable. Because of this, the theory of growth is largely a qualitative guide to historical studies, not a quantitative science.

Capital accumulation and technical change are big topics. It seems best to focus on a few aspects of how these factors have contributed to growth. Accordingly, this section looks at only three main topics: rising agricultural productivity, the growth of manufacturing, and the development of a national transportation network.

Productivity in agriculture

About 60 percent of the United States (excluding Alaska) is devoted to crop and pasture land. Another 15 percent or so is used for grazing. Judging from the pattern of *land use*, this is predominantly an agricultural country. Yet, fewer than one person in twenty lives on a farm nowadays. A century ago, the figure was closer to one person in two. Judging from the pattern of *employment*, we are no longer the nation of farmers we once were. But because so few people are farmers does not mean that this country isn't a major agricultural producer. Agricultural *output* in 1970 was six times what it was a century earlier, despite the fall in the farm population. The productivity increases in agriculture have been so spectacular that we can feed the entire U.S. population and still have a large surplus for export, using only about 3 percent of the labor force.

You can readily appreciate how important this rising agricultural productivity must have been to the general process of economic growth and development. The demand for farm products grows in rough proportion to the population. If agricultural productivity grows faster than the population, relatively fewer people are needed on the farm. With declining demand for farm labor and strong demand for labor in industry, the result is a migration from country to city, permitting industry to grow more rapidly than it could without a surplus farm population.

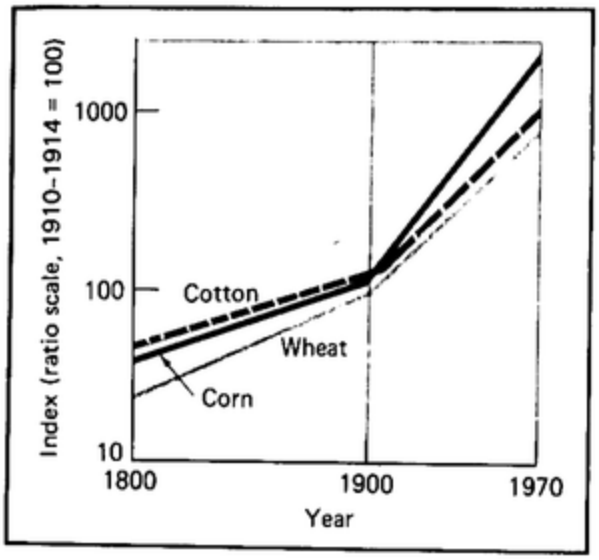


Figure 9 Output per hour of direct labor in agriculture 1800–1970

The growth in productivity of wheat, corn, and cotton was about 1 percent a year until early in this century, when it rose to about 4 percent a year.

Source: *Historical Statistics of the United States.*

Figure 9 shows the patterns of long-term productivity growth for three major crops—corn, wheat, and cotton. The patterns are broadly similar. Labor productivity from 1800 until the early 20th century grew about 1 percent a year. In the 20th century, however, the rate of growth jumped to about 4 percent. These magnitudes represent a fairly wide range of crops. Productivity in cattle and hog raising in the 20th century has not grown so fast, but in chicken raising, it has averaged about 6 percent a year. Nowadays, it takes about 1½ hours of direct labor to raise 100 broiler chickens to maturity. Chicken, which used to be a luxury item, is today one of the cheapest forms of animal protein.

These figures express productivity gain in terms of output per hour of work. Other productivity measures also show marked gains. From 1912 to 1970, gains in output per acre were about two or three times for wheat, corn, and cotton, and four or five times for hay and potatoes. Milk per cow and eggs per laying hen roughly doubled over the same period.

The factors responsible for this burst of productivity may be grouped under five headings:

1. Economies of scale.
2. The spread of irrigation.
3. The use of more and better machinery, fertilizer, and pesticides.
4. The development of better seed, animal feed, and animal breeding.
5. Improved techniques of cultivation and animal raising.

The economies of scale came not from increased use of land, but from consolidation of the land into larger farms. In 1900, 24 percent of farmland was used by farms of 1,000 acres (1.56 square miles) or more. By 1970, 54 percent of the land was incorporated into such large farms.

Irrigation is controversial, but about one third of the country's water is used to irrigate crops. Ninety percent of this is done in the West, where the water is transported over long distances at great expense to nourish land that would otherwise be ill suited to farming. An enormous capital investment was required to make this possible. Many people have argued that this investment would better have been diverted to some other use, that the opportunity cost of the irrigation exceeds its benefits. If so, irrigation has detracted from the growth in GNP, not added to it. But there is no doubt that it has added to agricultural development, particularly in the produce-growing areas of California.

A number of other factors contributing to the growth in agricultural output are summarized in Figure 10. There, you can see rapid growth in inputs of fertilizer, power machinery, and purchased inputs of feed, seed, and livestock. The contribution of more and better machinery is obvious,

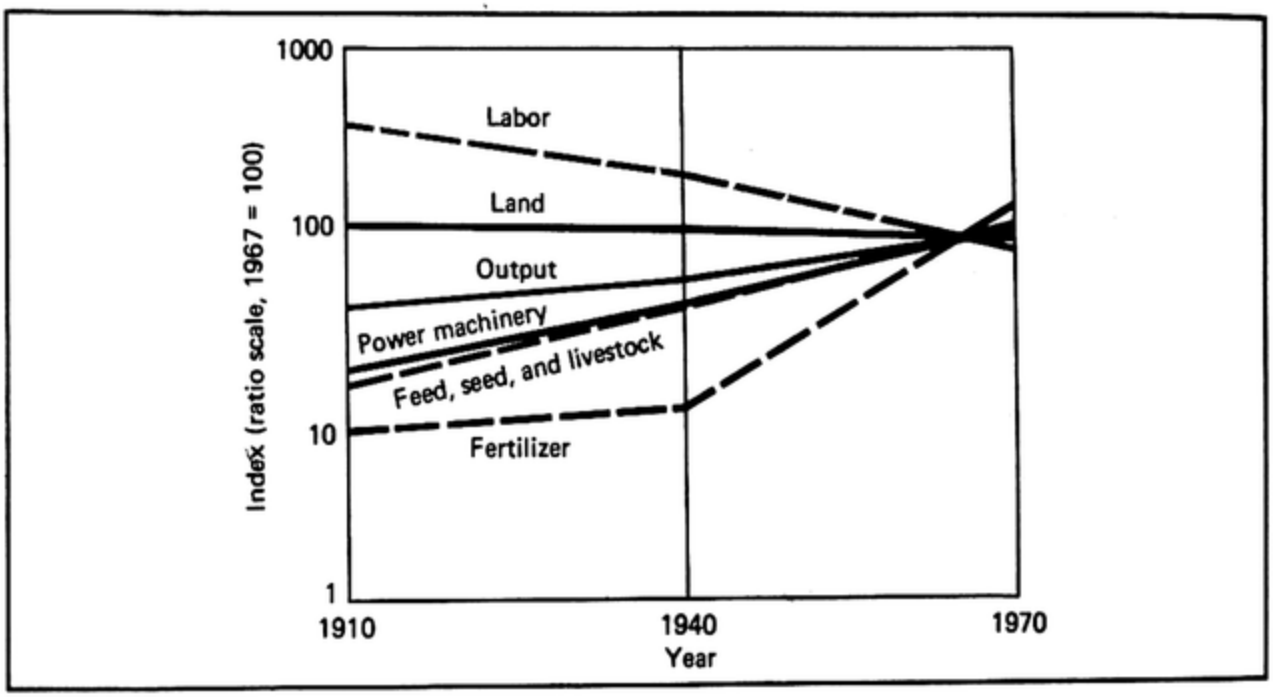


Figure 10 Farm inputs and output (1967–100)

Between 1910 and 1970, farm output increased by 130 percent, with little change in cultivated land. Labor input dropped by 70 percent, but the inputs of fertilizer, power machinery, and feed, seed, and livestock increased by 1780, 570, and 410 percent, respectively.

Source: *Historical Statistics of the United States.*

even to someone who has never set foot on a farm. This was the dominant source of growth in the late 19th and early 20th centuries. But you should not underestimate the importance of the growth of purchased fertilizer, feed, seeds, and livestock. At one time, farms were self-sufficient in supplying these inputs. Animals bred more animals, and supplied manure for the crops. Part of each year's harvest was put aside as seed for the following year. Animal feed was grown on the farm itself. Now fertilizers and feeds are produced in chemical plants, seeds are hybridized on special farms, and livestock are bred selectively on farms that specialize in raising young animals for sale. Farming is part of a very complex industry as closely tied to manufacturing by its inputs as it is by its outputs. Scientific farming has dominated the growth in agricultural productivity in recent decades. The result is appalling to people who romanticize family farming and organic vegetables. But it does provide nourishment to a large and still-growing population without absorbing a large proportion of the work force.

Because farm inputs are supplied from outside does not make farming easy: Modern scientific farming is a very exacting occupation, demanding as wide a range of skills as more traditional farming. But it is fair to say that much of the revolution in farm technology took place in the chemistry, agronomy, biology, and engineering laboratories of universities and manufacturing firms, rather than on the farm itself. Similarly, much of the capital investment that contributed to the growth in farm output took place in the farm input industry, not in the agricultural sector itself.

The growth of manufacturing

The verb "to manufacture" means literally to make by hand, and at one time, manufacturing was a handicraft occupation

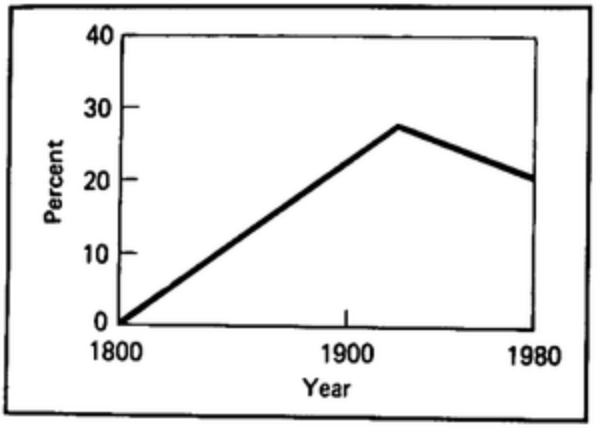


Figure 11 Employment in manufacturing as a percent of the labor force 1800–1980

In 1800, manufacturing employed only a small fraction of the labor force. It expanded rapidly during the 19th and early 20th centuries, reaching a peak of relative importance during the 1920s. Since then, it has lost ground relative to trade and service industries.

Source: *Historical Statistics of the United States.*

practiced in workshops and homes. This was largely true in this country during the Colonial period. Only the simplest sorts of manufactures were made in the colonies, which mainly produced food and raw materials and imported most manufactures from England. Thus, even though the colonies smelted pig iron, both wrought iron and steel were imported, as were even such simple iron products as nails.

In the 19th and early 20th centuries, however, the American economic landscape came to be increasingly dominated by manufacturing industries. This did not mean a spread of the old handicraft methods, but rather the development of the large-scale machine technology that is characteristic of 20th-century industry.

Figures 11 and 12 document some of the broad quantitative dimensions of this development. As you can see, manufacturing employment was negligible in 1800. By 1900, one worker in four was employed in manufacturing. Manufacturing employment as a share of the labor force reached a peak in the 1920s, but since that time, the fastest-growing industries have been

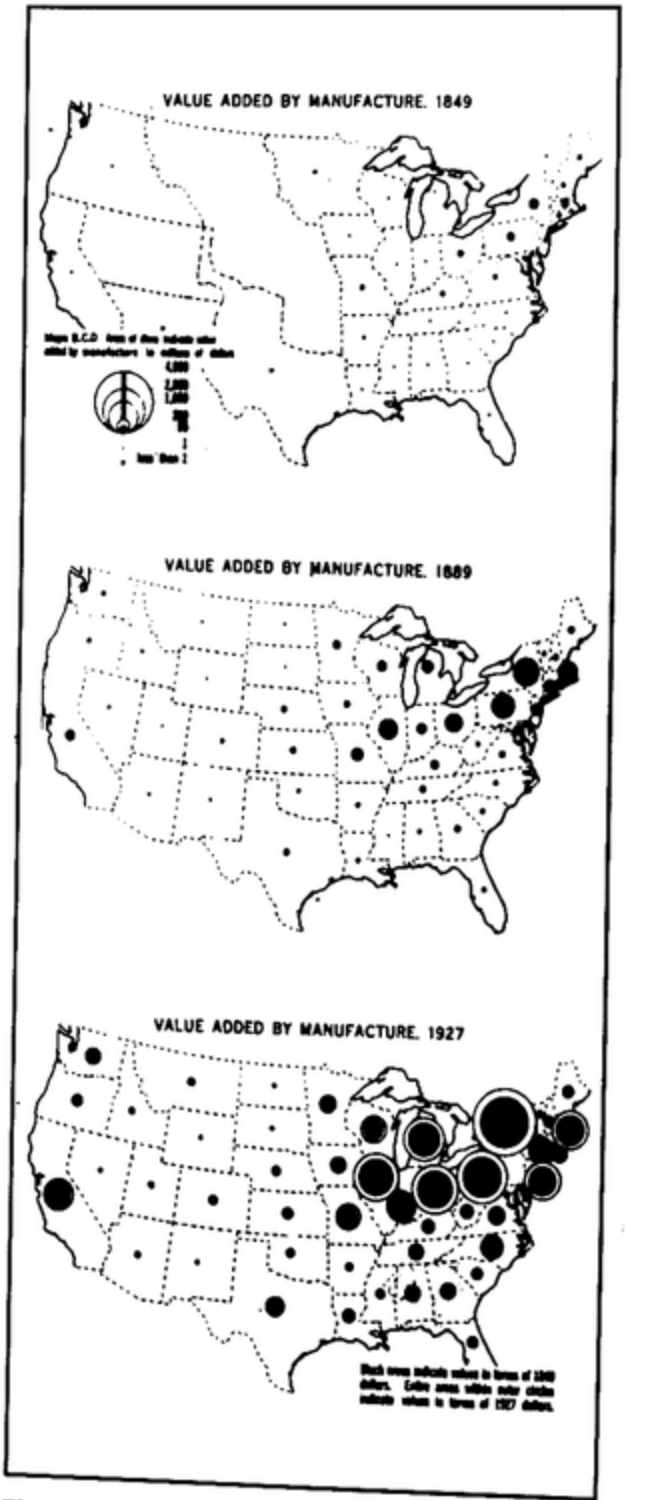


Figure 12 The growth of manufacturing across the country

The growth of manufacturing during the 19th and early 20th centuries was largely concentrated in the cities of the Northeast and Great Lakes regions, but by 1927, major manufacturing centers could be found in all of the 48 states.

Source: Paullin, *Atlas of the Historical Geography of the United States.*

trade, services, and government. Manufacturing is no longer the growth leader it was during an earlier period.

As Figure 12 shows, the growth in manufacturing was largely centered in the Northeast and Great Lakes regions. This was partly a reflection of the distribution of population at the time that industrialization began. It also reflected the dominance of the Southeast by slave agriculture. The wealth and the labor force of the South were locked into a mode of production that was less flexible than the free-labor, mobile-capital system of the North. A third factor of great importance was physical geography. Both the Northeast and the Great Lakes regions had coal and iron deposits and a system of natural waterways that overcame the substantial distances between centers of production and consumption. Thus, the growth of manufacturing in the North was largely the product of comparative advantage.

Some of the dimensions of the mechanization of manufacturing—the replacement of human motive power by water, steam, and electrical power—are fairly well documented from mid-19th century onward. The capacity of water and steam power sources used in manufacturing at midcentury was only about 1 million horsepower. Take the horsepower of the cars on the starting grid of the Indianapolis 500, multiply it a few dozen times, and you have it. After about 1850, this capacity increased tenfold, at a rate of about 5 percent a year (see Figure 13). The 20th century ushered in the age of the electric motor. Owing to the ease with which electrical energy is transferred from source to use, this form of power largely supplanted others and revolutionized machine technology.

Hard facts regarding the rate of capital accumulation aren't available until fairly late in the 19th century. But by 1879, accumulated capital in manufacturing amounted to about $35 billion in 1980 prices. This is a small figure by today's standards, but incomparably larger than what had been available in 1800.

As Figure 14 shows, about 75 percent of this accumulated capital was concentrated in what is usually called ***light industry***—food processing, textiles, forest products, and paper. These industries got their raw materials from the farms and forests, building on the natural resources that the

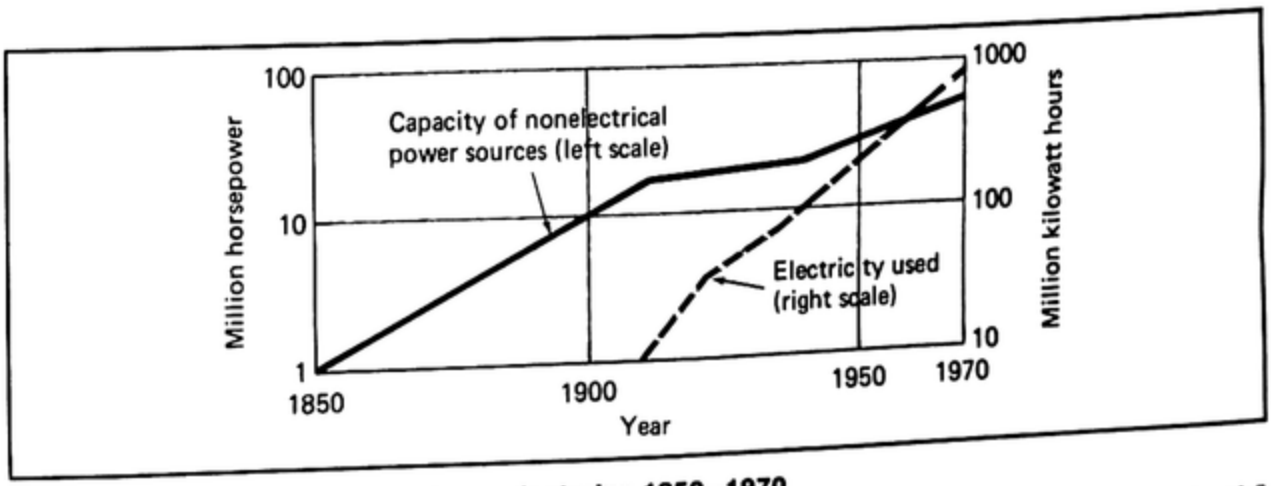


Figure 13 The mechanization of manufacturing 1850–1970

After 1850, the power capacity of engines (mostly steam) used in manufacturing increased at the rate of 5 percent a year. In the 20th century, the use of electricity in manufacturing increased by over 10 percent a year, revolutionizing production techniques in nearly every industry.

Source: *Historical Statistics of the United States.*

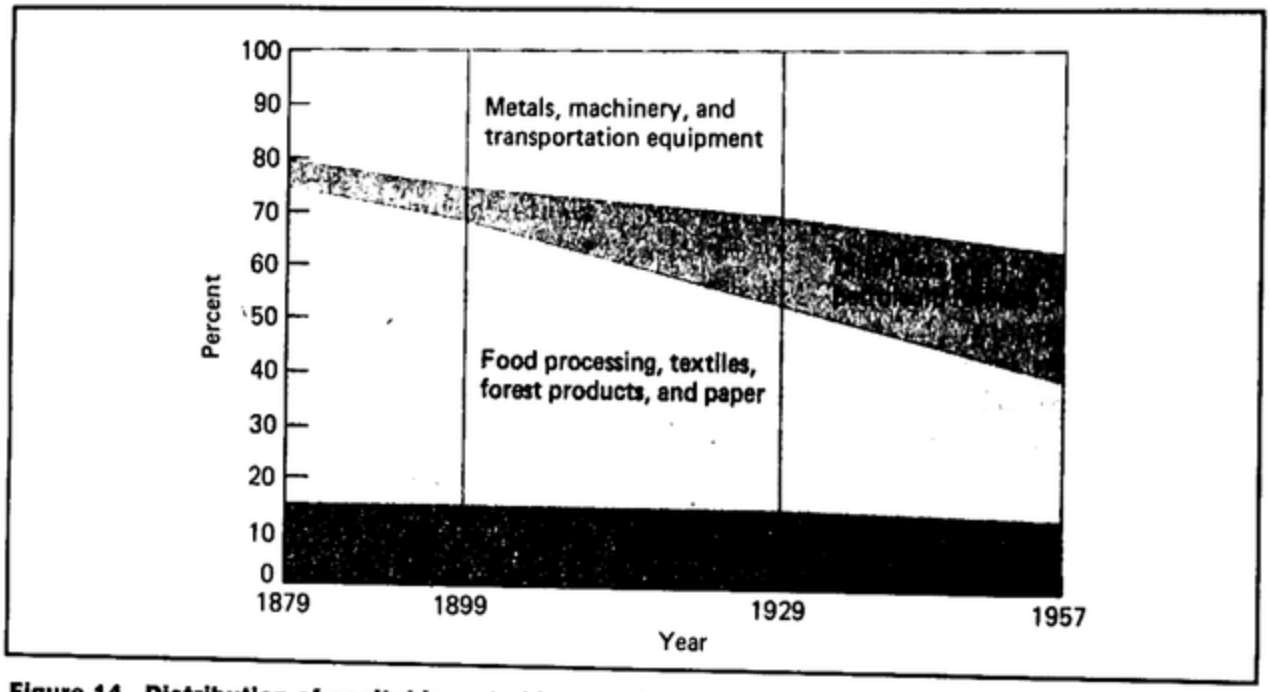


Figure 14 Distribution of capital invested in manufacturing 1879–1957
In 1879, capital in manufacturing was mainly invested in light industry—food processing, textiles, forest products, and paper. Over the succeeding 80 years, these industries declined in relative importance and were replaced by heavy industry—metals, machinery, transportation equipment, chemicals, and petroleum refining.
Source: *Historical Statistics of the United States.*

country had in such obvious plenty. The technology involved both in the manufacturing itself and in the provision of raw materials was relatively simple and ***labor intensive,*** so that a little capital produced a lot of product. At the end of the 19th century and throughout most of the 20th century, growth was largely centered in ***heavy industry***—primary and fabricated metals, machinery, transportation equipment, chemicals, and petroleum refining. These industries got their raw materials from mines and wells rather than from farms and forests. Both the manufacturing itself and the provision of its raw materials were technologically advanced and ***capital intensive,*** requiring a lot of investment before much product was forthcoming.

This pattern of light industry first, heavy industry second, is characteristic of the development of most of the industrialized world. But beneath these broad outlines, a far more complex pattern of interindustry relationships dictates what industries must develop in sequence, and what in parallel. To take an example, consider the interrelationships among the land-intensive agriculture of the Grain Belt, the railroads, and the metal and machinery industries. The development of midwestern agriculture in a sparsely settled region, far from its markets, required machinery and low-cost transportation. The railroad was the obvious solution to the transportation problem, but since railroads are not built without some prospect of markets, the development of agriculture and the railroad had to proceed simultaneously, not sequentially. Moreover, railroads cannot be built without machinery, equipment, and rails. These can be (and were) imported, but because they are heavy and costly to transport, the development of the railroad system depended

substantially on the development of the domestic locomotive, railroad car, and rail industries. These industries, in turn, depended on the development of primary steel capacity. The iron and steel industry used very heavy raw materials: iron ore, bituminous coal, and limestone. Its development, therefore, depended on the simultaneous development of the railroads. The whole mutually dependent system of agricultural, industrial, and transport development was delayed for a long time by the primitive technology of iron- and steelmaking. It was not until the invention of the Bessemer process for steelmaking in the 1870s and the open-hearth process in the early 1900s that durable rails could be produced at low cost. And it was not until the perfection of the internal combustion engine and the farm tractor in the 1910–1930 period that the potentialities of mechanized, land-intensive agriculture in the Grain Belt could be realized.

The complexities of the input-output structure create a maze of ***forward and backward linkages*** between one industry and both its customers and its suppliers. In a nicely calculated pattern of ***balanced growth,*** the various industries would expand at exactly the right relative rates, so that neither shortages nor excess capacity interfered with growth. It does not happen this way, of course. Both bottlenecks and surpluses develop. This is true in both planned and market economies. Some students of economic growth think that such ***imbalance*** is a creative force in development. A railroad with excess capacity creates a possibility for industrial and agricultural growth along its right-of-way. The owners of the railroad, whether they be socialists in Russia or capitalists in the United States, have an incentive to promote this growth. Similarly, a bottleneck in the interindustry flow of goods creates an incentive for capital accumulation and technological breakthrough. Because these tensions are only resolved with a lag, American economic growth has advanced in a pattern of alternating intensity and quietude, sometimes known as the "long swing" pattern. These episodes in capital accumulation and technical change are mirrored in corresponding periods of prolonged prosperity and slack. Some believe that the troubles of the 1970s and early 1980s reflect the beginning of another long downswing. If so, the next decade may not be particularly prosperous. But it is at least arguable that the creative tensions of the developmental booms produce a higher long-term growth rate than we would have if growth were more balanced and tranquil.

Transportation

If you have ever tried to hitchhike across the United States, you know something about how enormous it is. Along with this vastness comes great geographic diversity and the potential for regional specialization. But to realize this potential, a transportation network, uniting the specialized regions into a single national market, had to be built. Because of the distances involved and the difficulties in traversing both the eastern and western mountains, the building of the transportation network took more than a century after the American Revolution.

Nearly every transportation system requires enormous capital investment long before a big payoff is realized. A railroad built to join the eastern centers of population and manufacturing with the midwestern agricultural regions has great potential for raising national output and profit. But until it is actually completed, all the way from the East to the Midwest, over mountains and rivers, it is of limited value relative to its potential. Hence, private capital is characteristically slow to move into the transportation business without some

form of government assistance that makes it attractive to invest despite the long delays and risks in realizing a payoff. Both industry and agriculture can be "protected" against foreign competition by tariffs and other trade restrictions, but most forms of transportation require *direct subsidy* if they are to develop rapidly. Some of the most interesting chapters in American political history document the changing attitudes toward the subsidization of the transportation industry. The long-term trend has been toward greater public involvement, culminating in federal construction of the Interstate Highway System and the federal takeover of the bankrupt Penn Central Railroad in the 1960s and 1970s.

The earliest national transportation system, largely built in Colonial times, was the series of east-west wagon roads joining the agricultural area immediately west of the Appalachian Mountains with the eastern ports and population centers. You can see from Figure 15 that this was not a very extensive or well-integrated system, but then neither was the country well integrated in those days. The road network was largely built by private investors. Some consistency between the pieces of the road system was achieved by state governments, which licensed road construction. The builders of the roads received their revenues from charging tolls. We get our word "turnpike" from the series of poles, or pikes, stretched across the roads at periodic intervals. The user of the road had to pay a toll before the toll taker would turn the pike aside to let animals and wagons past.

The road system was not very good. Because there was no economically feasible way of "hardening" the roads with stone ballast and gravel surfacing, they were very crude, barely usable in wet weather, and particularly unsuited to carrying heavy loads. As you can appreciate, this greatly limited their usefulness in sustaining a geographic division of labor.

The second major transportation system, which complemented the roads rather than replaced them, was a system of canals. If you compare Figure 16 with Figure 15, you will see that the canal system covered some of the same territory as the roads. It was better suited to carrying heavy materials and foodstuffs, even though it was agonizingly slow. The canal barges were not self-powered but were drawn by animals. If you are ever in Washington, D.C., on a nice spring day, be sure to go for a jog or bike ride along the old towpath flanking the Chesapeake and Ohio Canal. You will see how narrow the canal is, and will be able to imagine how primitive the whole process was. Some of the canals were discontinuous at the most difficult parts of the eastern mountains. The barges were hauled out of the water and carried by rail over the high spots.

The only canal that was a real financial success was the Erie Canal in New York State. Like most canals, it followed riverbeds wherever possible. The route was particularly fortunate, since the Hudson River carried traffic between New York and Albany without the need for a canal. From Albany, the Mohawk led far inland with minimal dredging. And the total rise, from Albany to the highest point on the canal, was only about 600 feet, so that comparatively few locks had to be built. Because the outlay to construct the canal was low relative to the potential revenue from joining New York City to the Great Lakes, the canal company was able to withstand the financial panics and railroad competition that bankrupted most other canals. The Erie Canal still exists as the New York State Barge Canal. If you drive west along the Mohawk out of the Albany-Troy-Schenectady area, you can see it in use. Nearly all the others are in ruins and have been for a century. Most had

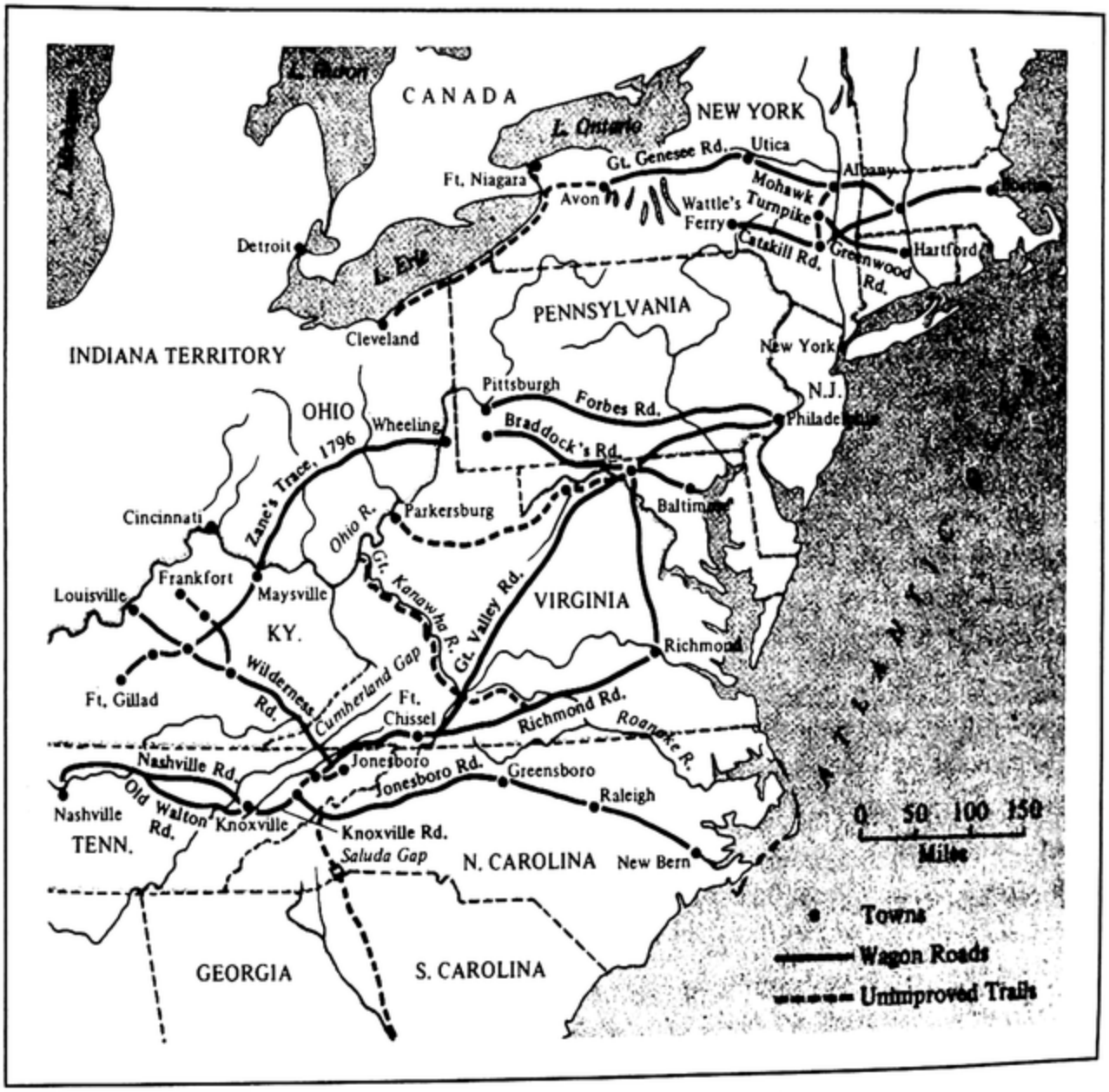


Figure 15 Main east-west roads in the first decades of the 19th century

The nation's first long-distance transportation system was a series of east-west roads joining the eastern ports and population centers with the agricultural region immediately to the west of the Appalachian Mountains. It was built almost entirely with private capital.

Source: W. Elliot Brownlee, *Dynamics of Ascent: A History of the American Economy*, First Edition. Alfred A. Knopf, Inc., 1974.

been extensively subsidized by state government bond issues, and the states shared in their losses. This greatly dampened the states' enthusiasm for subsequent subsidies to the transporation industry.

The steam engine transformed 19th-century transportation, as well as manufacturing. The steam engine made upstream navigation on the Mississippi, Missouri, and Ohio rivers and their tributaries possible. This increased the importance of natural inland waterways relative to canals, permitting shippers to move heavy cargo to and from New Orleans in relatively large vessels. The other thing it made possible was the railroad.

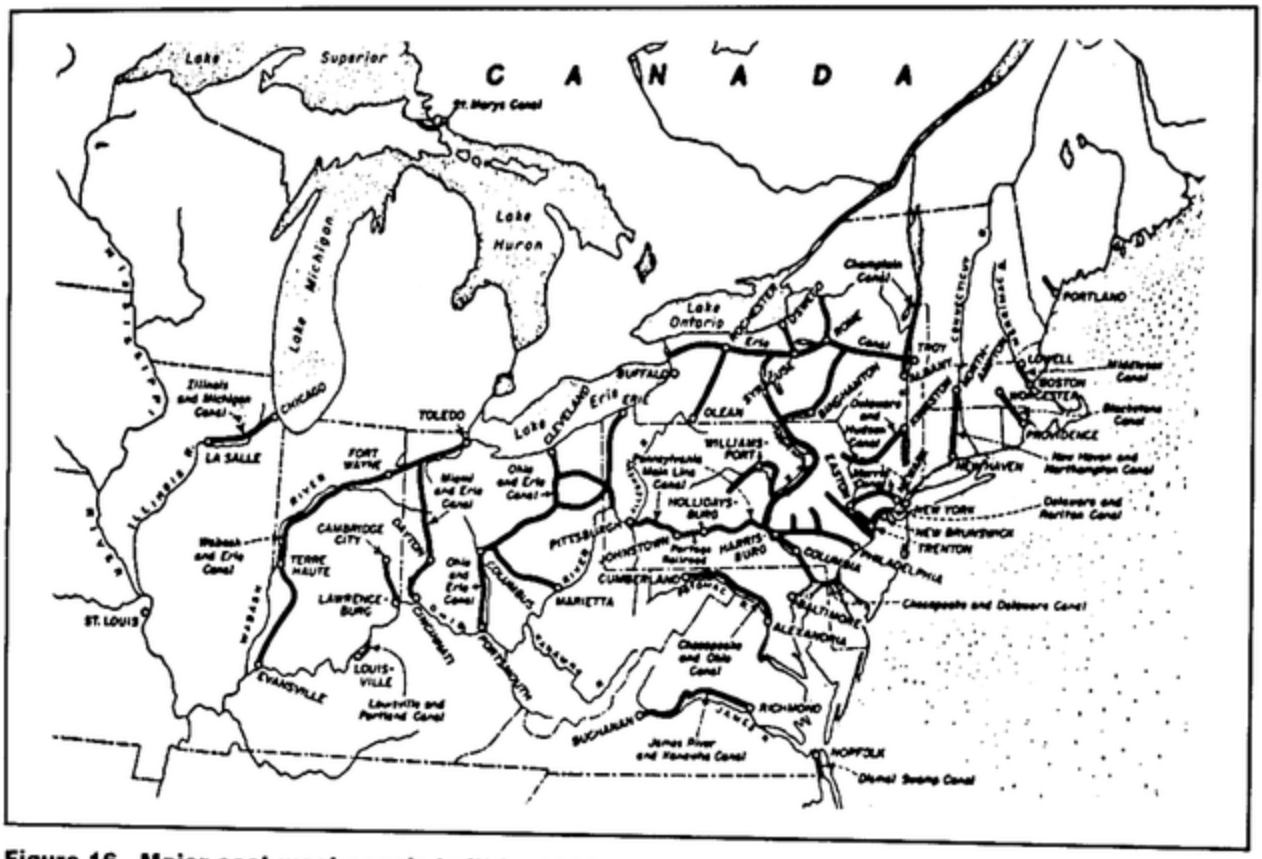


Figure 16 Major east-west canals built by 1860

The canal system was the second major transportation network. The states through which the canals passed helped private capitalists to finance their construction. When most of the canal companies failed in midcentury, the state governments shared in their losses.

Source: Brownlee, *Dynamics of Ascent: A History of the American Economy.*

Look at Figure 17. In Panel I, you can see the railroad system as it existed at the time of the Civil War. Notice how strangely disconnected it was, made up in part of short pieces of line joining nearby cities. This was particularly true in the South, but was also true to some extent in the rest of the country. Local entrepreneurs built lines to serve local needs—joining a mine to a smelting plant, a forest to a sawmill, or a plantation to a steamboat dock, for example. But the way to construct a railway that serves a regionally specialized national economy is to locate large, well-constructed *trunk lines* between major industrial, commercial, and population centers, with *feeder lines* connecting them to the smaller centers. You can see such a line stretching down the coast from Maine to North Carolina, and others developing to link the Midwest with the East. The South was not well served by rail lines in 1860 (although it did have a good network of waterways). This was one of the reasons it lost the Civil War.

Panels II and III of Figure 17 show what happened in the last 40 years of the century. By 1870, the Union Pacific linked San Francisco with the East. By 1900, the nation was a spider web of trunk and feeder lines. This rapid development of the railroad system dominated economic growth in the last third of the century. Without it, the development of the West would have been painfully slow, since the West was not linked to the East by inland

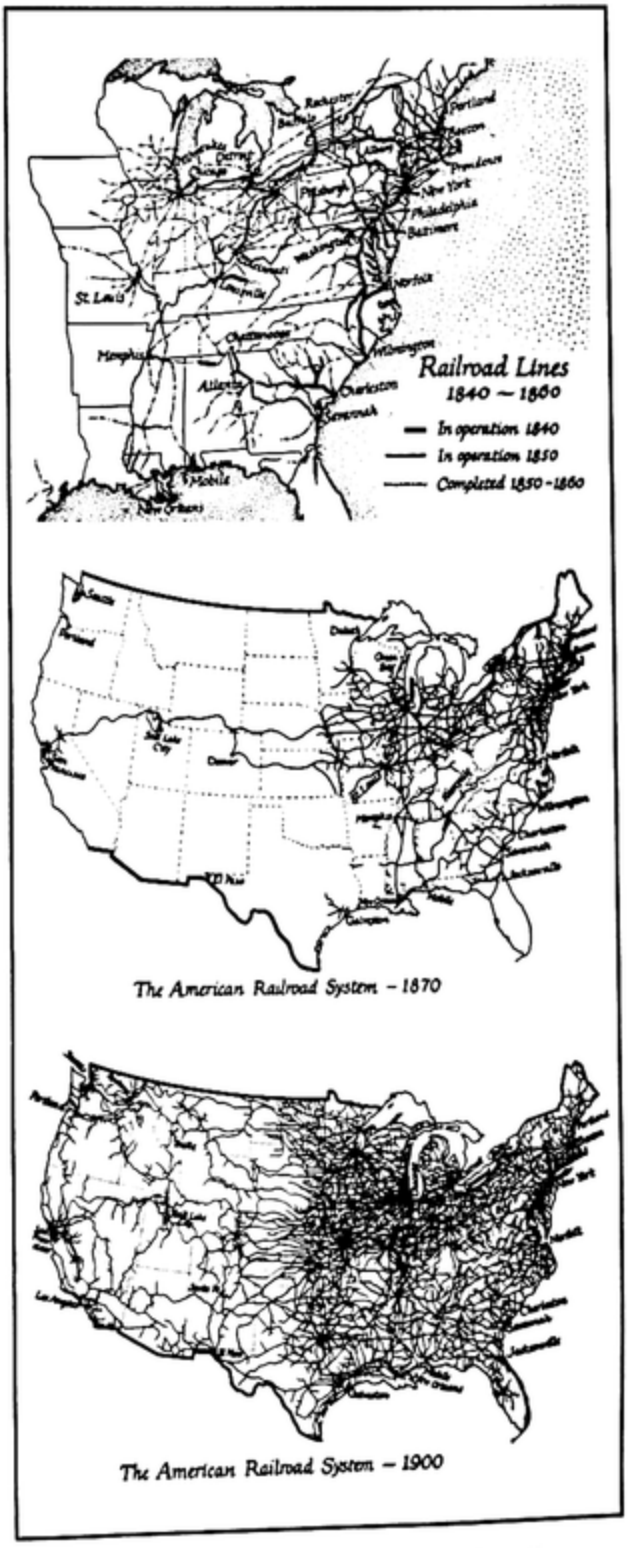


Figure 17 Railroad development after the Civil War

Between the Civil War and 1900, the American railroad network developed from a regional system to a trunk- and feeder-line system that served an increasingly specialized national economy.

Source: From Gilbert C. Fite and Jim E. Reese, *An Economic History of the United States*. (Boston: Houghton Mifflin Co., 1st ed., 1959) pp. 186 and 312. Used by permission.

waterways, and ocean traffic between East and West was slow and expensive until the Panama Canal was finished in 1914.

Railroad construction received a sizable federal subsidy. From 1823 through 1871, about 131 million acres of federal land, or about 7 percent of the area of the 48 states, was granted to the states to encourage railroad development. The states, in turn, granted the land to the railroad companies. Very little of it was actually turned into right-of-way or switching yards. Most was sold by the railroads to farmers, loggers, mining companies, and speculators, to finance the building of the railroads. The buyers developed the land and became customers for the railroads. This particular form of subsidy provided not only assistance in financing the railroad construction, but also the forward linkages that ultimately became a major source of operating revenues for the railroads.

The final links in the national transportation system—the highways and airports, along with the cars, trucks, and airplanes that use them—were built well into the 20th century. By this time, the country's industrial and geographical structure was similar in broad outlines to that of the country you are familiar with today. Neither the highway nor the airway system was responsible for populating the continent or turning it into a dual, agricultural-industrial country. There is no denying that the speed and convenience of these two modern modes of transportation have shaped the details of late-20th century economic life. But the outlines were already drawn 80 years ago.

What is truly distinctive about the highway and air transport networks is the way they were financed. The first system of roads in the country was financed with private capital. The building of canals was privately financed also, although with considerable assistance from the states. The

railroad companies got an indirect subsidy from the federal government, in the form of land grants, and a lot of subsidies from localities along their rights-of-way. But the highway and airport systems of the 20th century were built directly with federal, state, and local government funds. Only the carriers that used them—the cars, trucks, and planes—had to be financed with private funds. In building the highway and airport systems, the various levels of government recognized explicitly that the development of a national transportation system poses distinct financing problems because of the long delays between investment and its payoff. Rather than waiting for private capital to respond to subsidies, the governments attacked the problem directly. In doing so, they probably set a pattern for the future.

Summary

The study of American economic growth is a major undertaking, and it is embedded in an even bigger undertaking, the study of economic history. This chapter described only a few of the major trends that transformed this country from a nation of small farmers and artisans into a modern agricultural and industrial power.

1. Economic development in the 19th and early 20th centuries was fueled by massive immigration from Europe. During some periods, immigration contributed more to population growth than did natural increase. Immigration is superior to natural increase as a source of labor force growth because adult immigrants arrive ready to work, and do not have to be nurtured and educated before they can begin their economic lives.
2. Internal migration played a major role in populating and developing the Great Plains and the West. As European immigrants were moving into the East and Midwest, native-born whites were moving westward in search of cheap land and high wages.
3. In the 20th century, large numbers of blacks moved from the South to the industrial centers of the East and Midwest, making possible a rapid development of industry during World War II and the postwar prosperity.
4. The growth and migration of the population were major contributors to economic development, but equally important was the upgrading of the productive population in ways that made it more suitable to the needs of the market economy. Among the major elements in this upgrading were greater participation in the market labor force, better health, and more schooling.
5. Population growth cannot in itself sustain economic development for long because of the problem of diminishing returns. It must be accompanied by capital accumulation and technical change, forces that are closely interrelated. Three of the major dimensions of capital accumulation and technical change in American economic development were the scientific revolution in agriculture, the development of a mechanized, high-technology manufacturing industry, and the building of a national transportation network.
6. The scientific revolution in agriculture proceeded on two fronts. Farming practices themselves changed because of the increased use of machinery, irrigation, and manufactured fertilizer and pesticides, and because of the introduction of genetically improved plants and animals. But these developments on the farm required the parallel development of supplying indus-

tries to manufacture tractors and fertilizers and to breed better plant and animal strains. As a result of the agricultural revolution, the share of the labor force engaged in farming dropped from about 50 percent to about 3 percent between 1870 and 1970, despite a sixfold increase in output.

7. The growth in manufacturing, like that in agriculture, was a process of mechanization and application of scientific method. Early manufacturing was mainly light industry, using raw materials from the farms and forests in labor-intensive processes. This was succeeded by growth in heavy industry, applying high-technology, capital-intensive techniques. The high-water mark for manufacturing was reached in the late 1920s, when nearly 30 percent of the labor force was engaged in manufacturing. One of the most important social by-products of the growth in manufacturing was the urbanization of much of the country, especially the East and the Midwest.
8. The development of the transportation network proceeded in waves over two centuries, determined in large part by improvements in transport technology. First came roads, then canals, railroads, highways, and finally airports. An important influence on the pace of transportation development was the trend toward greater government participation in the financing of capital formation.
9. Developments in agriculture, manufacturing, and transportation were not independent, but interrelated by the linkages that make up the input-output system. Sometimes these three sectors developed in balance with one another, and sometimes their imbalances created tensions that led to major developmental breakthroughs.

Key concepts

Demographic changes, capital accumulation, technical change
Immigration
Natural increase
Dependency ratio
Internal migration
Urbanization
Light and heavy industry
Labor and capital intensive
Forward and backward linkages
Balanced versus imbalanced growth

Questions for review

1. a. Why is population growth due to immigration more conducive to rapid economic growth than population growth achieved through natural increases?
   b. If immigration is conducive to rapid economic growth, why were severe restrictions placed on immigration in the 1920s?
2. Choose four of the demographic factors influencing the rate of economic growth that are depicted in Figure 7. Explain carefully how each of these four factors influenced the rate of economic growth.
3. Describe some of the factors that would cause a fall in a country's dependency ratio.
4. Explain why technological improvements are so important if a country is to experience sustained economic growth.
5. Industrialization and urbanization are usually considered the cornerstones of economic growth and development. Why, then, can it be said that rising agricultural productivity was a crucial factor in U.S. economic growth?

34

International Trade

As you read and study this chapter, you will learn:

- why countries can benefit from international specialization and world trade
- how world demand influences the distribution among countries of the gains from trade
- what the major arguments are for restricting trade, and how tariffs and quotas work
- what the pattern of American trade looks like

Even if you knew everything about producing goods and services and had access to all the world's natural and capital resources, you could hardly produce on your own the quantity and variety of goods that make up your ordinary pattern of consumption. Like everyone else, you benefit from cooperative production, specialization, and the division of labor.

As Adam Smith taught us, "The division of labor is limited by the extent of the market." Big markets enable people to specialize very narrowly. Without large cities to support symphony orchestras, there would be no professional piccolo players. Today's *world market* is simply the division of labor and specialization carried about as far as it can go. Yet, the American market is as large and complex as the world market was a few generations ago. Why do we trade beyond the boundaries of such a large national market? Don't the gains from specialization stop once the market reaches a certain size? Perhaps a system of indepen-

dent national economies would be as prosperous as one made up of international trading partners.

It is quite obvious that countries benefit from trading for products they cannot produce at all. If Switzerland has no iron ore, it must import ore, steel, or golf clubs if the Swiss want to play golf on their home soil. It is also fairly obvious why Iceland imports coconuts, even though it probably could grow them in hothouses warmed by hot springs. But it is less obvious why industrial countries trade so intensively with one another, and why agricultural countries specialize in a few products for the world market while they import other agricultural products not much different from those they grow.

Comparative advantage and the benefits of trade

If all countries had the same resources and were at the same stage of economic development, there would be few gains from trade. But whenever countries differ in natural, human, or capital resources, each benefits from producing those things that are best suited to the resources it has. David Ricardo first demonstrated this principle in the early 19th century. It is known as the ***law of comparative advantage.***

Comparative advantage

Ricardo's proposition is easiest to understand if you start with a very simple example. Suppose the world is made up of only two countries, Holland and Belgium. They are both capable of producing only two products, wheat and tulips, which require only land and labor for their production. Belgium and Holland have equal population and land area, but because Belgium's soil is poorer than Holland's, its agriculture is less productive, particularly in the cultivation of tulips. Within each country, land and labor can be shifted back and forth between tulip and wheat production without a change in the relative costs of the two outputs. Finally, there are no transport costs, either internally or between countries.

Such simple assumptions underlie the two *production possibility curves* shown in Figure 1. Each curve gives the outer boundary of the outputs that its economy can produce—the *maximum* amount of one good for each given amount of the other.

According to Figure 1, Belgium can produce at most 3 carloads of wheat per week, and at most 1 carload of tulips. Since resources can be shifted from one crop to the other without changing relative costs, the production possibility curve is a straight line. Competitive pricing will make tulips three times as expensive as wheat without foreign trade.

Holland is more productive than Belgium in both wheat and tulips. It can produce 4 carloads of wheat a week compared with 3 for Belgium, or 3 carloads of tulips compared to 1. This means that Holland can produce either good using fewer resources per unit of output than Belgium uses. Holland is said to have an ***absolute advantage*** over Belgium in both wheat and tulips. Its production possibility curve lies entirely outside Belgium's, so Holland is clearly the richer country.

But this is not the crucial difference between the two. What is important for Ricardo's argument is that Holland's productivity advantage over Belgium is greater for tulips than it is for wheat. In Belgium, 1 unit of wheat buys $^1/_3$ of a unit of tulips without trade. In Holland, 1 unit of wheat buys $^3/_4$ of a unit of tulips. Tulips are relatively cheap in Holland and wheat is relatively cheap in Belgium, in terms of

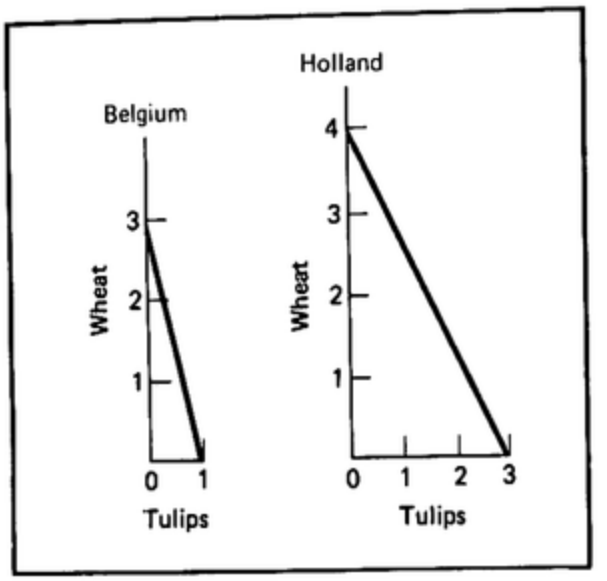


Figure 1 Production possibilities in Holland and Belgium

In Belgium, the labor force can produce 1 carload of tulips a week if it produces no wheat. For every unit reduction in tulip production, it can increase wheat production by 3 units. This means that Belgium can produce 3 carloads of wheat a week if it produces no tulips. In Holland, where the land is more fertile, the labor force can produce 3 carloads of tulips a week or 4 carloads of wheat. However, its advantage over Belgium in wheat production is less great than its advantage in tulip production. When resources are shifted from tulip production to wheat production, wheat output only rises by ⅓ of a unit for every unit reduction in tulip output. Its maximum possible output of wheat is only a little larger than that of Belgium. Thus, although Holland has an absolute advantage over Belgium in both wheat and tulips, Belgium has a comparative advantage in wheat. In other words, wheat is comparatively cheap in terms of tulips in Belgium and comparatively expensive in Holland.

opportunity cost. Holland has a *comparative advantage* in tulip production, and Belgium has a comparative advantage in wheat production. The Dutch can get more wheat for their tulips by trading with the Belgians than they can by shifting their domestic resources from tulip production to wheat production. And since it costs *relatively* little to produce wheat in Belgium, the Belgians can get more tulips for their wheat by trading with the Dutch than they can by shifting resources at home. It sounds as though they ought to be able to work a deal, doesn't it?

Specialization and gains from trade

There is room for a deal because there are ***gains from specialization and trade.*** To see what these benefits amount to, it will help to look at a production possibility curve for Holland and Belgium combined, a "world" boundary for production. This is shown in Figure 2. It is constructed from the production possibilities of the two separate countries, assuming a pattern of efficient specialization.

First, suppose that both countries specialize entirely in wheat output, producing no tulips at all. The result is 7 carloads of wheat a week, 3 from Belgium and 4 from Holland, at Point *A*. Now suppose that some tulips are to be produced. Efficiency requires that they be produced in Holland, the country that can most efficiently shift its resources to transform wheat into tulips (the production possibility curve is also called the *transformation curve*). Any world tulip output up to 3 carloads a week should be grown entirely in Holland. Doing otherwise means going to Belgium's high-cost tulips when the low-cost source can still produce more tulips. But once Holland's entire economy is specialized in tulip production, at Point *B*, the low-cost source of tulips dries up. Then, additional tulips can only come from Belgium. The rate of transformation from wheat into tulips takes a sudden turn for the worse, since Belgium is the high-cost producer.

If you look carefully at the world production possibility curve, you will see that it is pieced together from the separate curves for the two countries. Triangle *ABC* is Holland's curve; *BDE* is Belgium's. From *A* to *B*, Belgium's wheat production is constant at 3, and its tulip production is zero. Only Holland's output is changing. Between *B* and *D*, Holland's tulip output is constant at 3, and its wheat output is zero. Only Belgium's output is changing.

Note also that at every point of efficient world production, *at least one coun-*

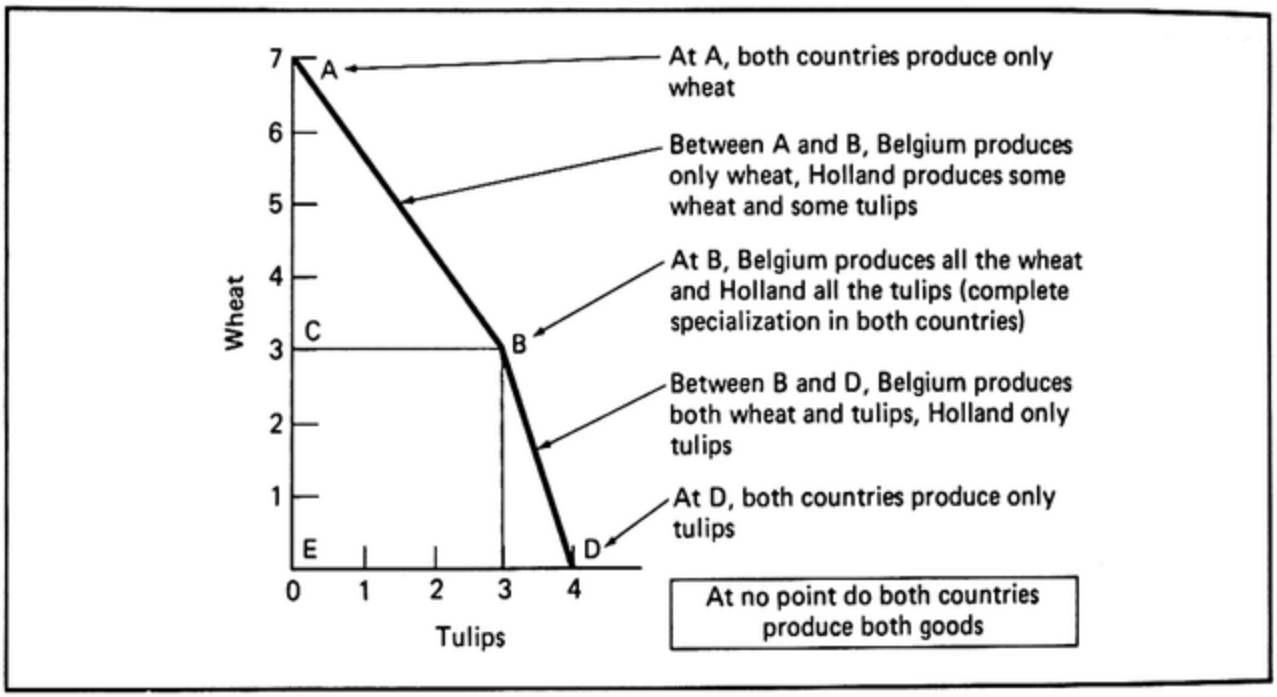


Figure 2 The world production possibility curve
Efficient production requires that at least one country specialize completely.

try is completely specialized. That is, nowhere do both countries produce both goods. This is characteristic of goods produced under constant costs. If relative costs are constant (and different) in two countries, it cannot be efficient to produce both goods in both places.

Figure 3 shows two patterns of inefficient production. Suppose that Belgium is completely specialized in wheat and Holland produces both goods. What will happen if Belgium then produces some tulips? It can only do this by moving down *its own* production possibility curve. This must move the world down *FG*, for example, on Belgium's terms, rather than down *FB*, on Holland's. This carries the world into the inefficient region, inside its production possibility curve. Similarly, if the world is at *H*, and Belgium is not entirely specialized in wheat production, it is inefficient for Holland to produce wheat. This would mean moving along *HI*, for example, into inefficient territory. Better to produce the additional wheat in Belgium.

Now think about one final matter. Suppose that Belgium can produce wheat and tulips in the ratio 3 to 1, as before, and that Holland produces them in the ratio 6 to 2. That is, Holland can produce twice as much of both goods in any given proportion and has an absolute advantage in the production of both. *But now no country has a comparative advantage in either good.* If you were to try to reconstruct Figure 2 on these assumptions, you would find that it has no kink. The rate at which wheat can be converted into tulips by shifting resources is the same in both countries. *Because of this, it doesn't matter which country produces which good.* It follows that there is no advantage from specializing, and there are no gains from trade. Gains only come from exploiting comparative advantage.

You should not imagine that gains from trade are equally shared. The distribution of gains depends on demand. In the Holland-Belgium example, the distribution of gains depends on the demand for

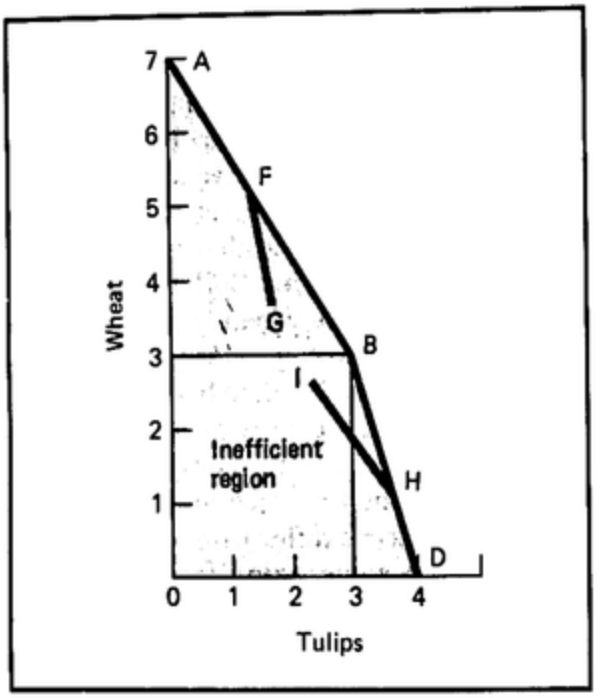


Figure 3 Inefficient production

If Belgium is completely specialized in wheat and Holland is not completely specialized in tulips (say, at Point *F*), then producing some tulips in Belgium means moving along *FG*. But every point on *FG* is *inside* the production possibility curve and is, therefore, inefficient. Similarly, producing some wheat in Holland when Belgium is incompletely specialized (at *H*) means moving along *HI*, into an inefficient pattern of world production. No country that has a *comparative* disadvantage in a good should ever produce it unless the country that has a comparative advantage in that good is already completely specialized in its production.

wheat relative to the demand for tulips. If the demand for wheat is very strong, Belgium, the country with the comparative advantage in wheat, will make the biggest gains. This is because the high demand for wheat pushes production into the *AB* portion of the production possibility curve of Figure 2, along which prices reflect Holland's relatively high cost of wheat production. Since all wheat will command the same price, the country with the lower cost, Belgium, makes the bigger gains. But if the demand for wheat is weak relative to the demand for tulips, production will take place along the *BD* portion of the production possibility curve. Relative prices will reflect Belgium's relative costs, so that tulips will command a high price and Holland will make the larger gains from trade. Its comparative advantage in tulip production will make it prosper when tulip demand is relatively strong. In general, the countries that benefit the most from trade are the lowest-cost suppliers of goods whose world demand greatly exceeds their own production capacity.

The Ricardian argument for free trade is compelling. It was taken up by the British manufacturers in Ricardo's day. Armed with his ideas (and a lot of political clout), they persuaded Parliament to repeal the "Corn Laws," which were tariffs on imported grain. The resulting cheap grain imports crippled British agriculture, drove farm workers into the manufacturing cities, and lowered money wages relative to the prices of manufactured goods. Britain then specialized in manufacturing and became the dominant industrial power of the 19th century.

Beyond constant cost

The analysis you have just finished has ignored both transport costs and increasing or decreasing production costs. It is not too hard to see how to incorporate these cost factors into the theory of trade.

Transport costs are a barrier to trade, and even when gains from international specialization offset the costs of shipping goods from one country to another, shipping costs nonetheless reduce the gains from trade. In determining whether trade is beneficial, what matters is the *total* amount of domestic product that must be given up to pay for imports. Holland may be able to trade tulips to Belgium and end up with more goods than if it produced both wheat and tulips domestically. But if it has to pay the equivalent of a large amount of wheat to carry the wheat from Belgium, the *net* result of trade may be

fewer goods than it could get from producing both its wheat and tulips with domestic resources. If this is true, Holland is better off growing its own wheat and not specializing.

Transport costs introduce an important qualification to all arguments based on comparative *production* costs alone. If the cost of shipping between countries is high enough relative to differences in production-cost ratios, Belgium may be the cheapest source of both wheat and tulips for the Belgians, and Holland the cheapest source of both for the Dutch, even though their internal price ratios are different.

Transport costs also affect patterns of *internal* specialization. Bread and beer are produced all over the United States, even though there might be economies to *producing* them in just one place. Any such economies are offset by the transport cost of supplying them from a central source. And you should not be astonished to find a large or mountainous country exporting a good across one of its borders and importing the identical good across another border. Internal transport costs may well be prohibitive, while international transport costs are not. One coast is supplied by a low-cost, nearby foreign supplier, while the other supplies its nearby neighbors at least cost. The United States sells Alaskan oil to the Japanese while it imports Arabian oil through its East Coast ports.

Finally, consider the impact of returns to scale on gains from trade. If countries can take advantage of increasing returns and the resulting decrease in production costs, then the gains from trade will be even greater than if costs are constant. Producing for the world market will make it possible for a country to exploit its returns to scale. If tulips get progressively cheaper in terms of wheat as Holland specializes in tulips, and wheat gets progressively cheaper in terms of tulips as Belgium specializes in wheat, then the two countries can gain enormously from specializing. With decreasing costs, countries can benefit from trade even though they have identical technologies. By specializing, they create comparative advantage.

Increasing cost, however, is the enemy of specialization and trade. If the relative cost of tulips rises as Holland produces more tulips and less wheat, then Holland's comparative advantage shrinks as it becomes more specialized. Where increasing costs are the rule, countries will diversify rather than specialize completely. But whenever countries' technology or consumption patterns differ from one another, there will be some gains from trade. As long as relative costs differ between countries, it will pay them to redistribute output toward comparative advantage.

Many goods and many countries

Obviously, there are more than two countries in the world, and more than two commodities. The theory of trade with many countries and many goods cannot easily be presented by means of diagrams, but what you already know about comparative advantage is generally valid.

Suppose, for example, that besides Holland and Belgium, Denmark can also produce wheat and tulips and at a 1-to-1 cost ratio. A carload of tulips costs 3 carloads of wheat at Belgium's cost ratio, $^4/_3$ carloads of wheat at Holland's cost ratio, but only 1 carload of wheat at Denmark's cost ratio. Compared to both of the other countries, Denmark has a comparative advantage in tulip growing. Neither Holland nor Belgium should produce tulips unless Denmark is completely specialized. **In general, no two countries should ever produce the same two goods unless they have the same cost ratio between them, including transport costs. If they produce at different cost ratios, world production is inefficient.**

Now suppose that there are only two countries again, but three goods—wheat, tulips, and butter. Along Belgium's production possibility curve, its resources can produce 3 carloads of wheat, 1 carload of tulips, or 1 carload of butter a week. Along Holland's, its resources can produce 4 carloads of wheat, 3 carloads of tulips, or 4 carloads of butter a week. Who has the comparative advantage in what? The answer is that in the production of wheat, Belgium's resources are $^3/_4$ as productive as Holland's, in tulips they are $^1/_3$ as productive, and in butter only $^1/_4$ as productive. Belgium is at an absolute disadvantage in everything, but its disadvantage is least in wheat and greatest in butter. As long as Belgium can increase its wheat production, Holland should never produce wheat. For Holland to produce another carload of wheat costs the world $^3/_4$ carload of tulips or 1 carload of butter; in Belgium, the opportunity cost is only $^1/_3$ carload of tulips or $^1/_3$ carload of butter. Similarly, Belgium should never produce butter unless Holland is completely specialized in butter production. The order of comparative advantage from Belgium's viewpoint is wheat, tulips, and butter. From Holland's viewpoint, the order is just reversed.

These two generalizations to several countries and several goods have been explained in terms of the constant-cost analysis of Ricardo. But their lesson is quite general. **It doesn't pay for one country to make any good internally, along its production possibility curve, when the same good can be made in some other country at lower opportunity cost.**

The pattern of world trade

Because the world has many goods and many countries, economic life is more varied and interesting than the wheat-tulip life of Belgium and Holland. One dimension of this variation is the complexity of the pattern of world trade. Although the theory of comparative advantage is easiest to understand when it is reduced to barter between two countries trading two goods, such exchanges are rare in the world market. Instead, exporting firms sell their output to buyers in a variety of other countries. Each country's importers buy from a variety of exporting countries. There is no reason at all for the destinations of the exports to match up with the origins of the imports. Since trade is carried out by means of money and credit, accounts need not balance between any two countries. Country A may export only to B, C, and D, and import only from X, Y, and Z. There is no need for *bilateral* (two-sided) *exchange.* In principle, all exchange could be *multilateral*, involving many countries. In fact, it is a mixture of both.

You can get some insight into the pattern of ***multilateral trade*** by looking at Table 1, which shows exports among and within most of the world's major trading areas in 1979. The rows of the table show where the exports of the various countries go; the columns show imports. For instance, in 1979, North American countries exported $51 billion worth of goods to "Other Third World" countries (which are undeveloped countries other than the members of the OPEC oil cartel).

The table covers only merchandise trade, that is, trade in goods. There is no allowance for trade in services, such as transport or travel, nor are there entries for international income payments. Purely financial transactions, such as international lending and foreign aid, are also omitted.

There are a number of interesting things to note about this table:

1. There is considerable variation in the size of the numbers along the boxed diagonal relative to the row totals. Western European economies, which are

Table 1 The pattern of commodity trade among developed and Third World countries, 1979 (in billions of dollars)

	Destination of Exports					
	North America	Japan	Western Europe	OPEC	Other Third World	Total Exports*
North America	69	21	57	16	51	229
Japan	28	0	17	13	32	103
Western Europe	47	8	482	47	68	703
OPEC	47	33	75	3	44	207
Other Third World	52	23	56	11	40	199
Total Imports*	252	96	734	98	257	

NOTE: There is no need for the exports from one region to a second to be balanced by exports from the second to the first.
Source: GATT, *International Trade, 1979–80.*
*Details do not add to totals because of the omission of some areas, notably Eastern Europe, the Soviet Union, and China.

highly varied and diversified, trade heavily with one another. About 70 percent of their exports go to other countries within the same region. By contrast, the OPEC countries, all of which are heavily specialized in oil production, hardly trade at all with one another. Trade among North American countries is somewhere in between according to the figures. This is somewhat misleading. North America as a trading area comprises only two large countries, the United States and Canada. Much of what shows up as international trade among the smaller countries of Western Europe is internal trade for each of the two North American giants. The 30 percent of North American exports that go to North America reflect an enormous interdependence between the United States and Canada.

2. Because trade is partly multilateral, the figures that are symmetrically placed around the boxed diagonal don't have to match. In 1979, Japan exported $28 billion worth of goods to North America; North America exported only $21 billion to Japan. North America exported $57 billion to Western Europe, but only $47 billion went in the other direction.
3. Because commodity trade is only part of the pattern of international payments, a country or region may export more than it imports or vice versa. In 1979, OPEC sold over $100 billion more on world markets than it bought. The rest of the world went deeply into debt to buy its oil.
4. Despite the importance of oil and other primary products from the Third World, trade among the developed countries of North America, Japan, and Western Europe dominates world trade. About half of world exports flow from one developed country to another.

Protectionism

Despite the force of arguments based on the theory of comparative advantage, most governments interfere with the flow of free international trade. The most common form of interference is ***protectionism***, a deliberate policy of helping domestic in-

dustries meet import competition. There are many arguments for protectionism. A few are good, some look sound but are faulty, and others are transparently false.

You can easily see, though, how American industries might persuade the government to protect them from competition. U.S. corporations are very involved in domestic politics. The President and Congress are well aware of industry's interests, and lobbyists spend millions to promote this awareness. The Americans who are hurt by protectionism are barely aware of what it costs them. What is amazing is not the continued existence of protection, but the extent to which trade has become progressively freer. Tariffs, which are taxes on imports, are only about 20 percent as large as they were a century ago.

This part of the chapter explains how most common forms of protection work and then surveys some of the arguments in favor of protectionism. After you have analyzed these arguments and compared them to those of the previous section, you ought to be able to make an independent judgment about the relative desirability of free trade and protection.

Tariffs and quotas

A ***tariff*** is a tax on imports. It raises the prices of foreign goods relative to domestic substitutes. In earlier centuries, tariffs performed two functions. First, they increased the incomes of domestic suppliers of the taxed goods by raising prices. Second, they provided a major source of government revenue. Governments used to be a lot less effective in collecting domestic taxes than they are now. Their ability to control major seaports made tariffs very attractive because they were easy to collect. Today, tariffs have been largely supplanted as a revenue source by income, sales, property, and value-added taxes, and the main function of tariffs is to restrict foreign competition.

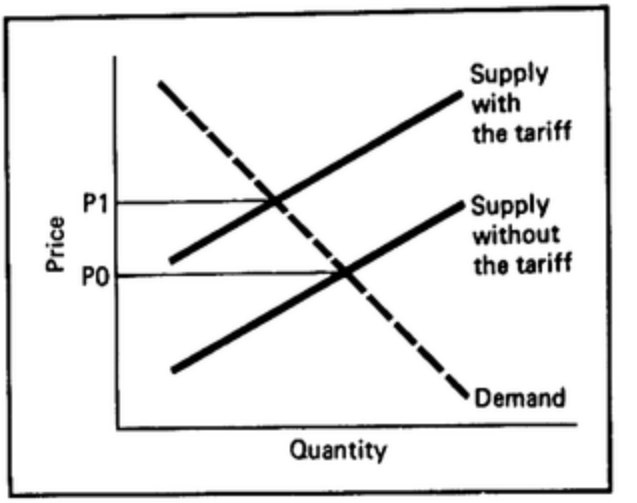


Figure 4 A tariff

If a tariff is imposed on the importation of a good, the supply curve shifts up. Price rises and quantity drops.

To grasp how a tariff works, picture a supply and demand diagram like Figure 4. Then suppose a tariff is levied on a *part* of the supply. The supply curve shifts up. Quantity supplied declines, and price rises. Those suppliers who are not taxed benefit. They are domestic producers. Those who are taxed suffer. They are foreign producers. Domestic consumers pay more and get less.

A ***quota*** is a legal restriction on the quantity of a good that may be imported. The quota may be a total ban on imports or a partial restriction. Like a tariff, it shifts the supply curve for a good upward. But the two forms of supply restriction operate differently. A quota directly reduces quantity supplied and indirectly raises price. A tariff directly raises price and indirectly reduces quantity supplied. In other respects, however, their effects are equivalent. To see in greater detail how quotas and tariffs work, study the accompanying box. It stresses both similarities and differences.

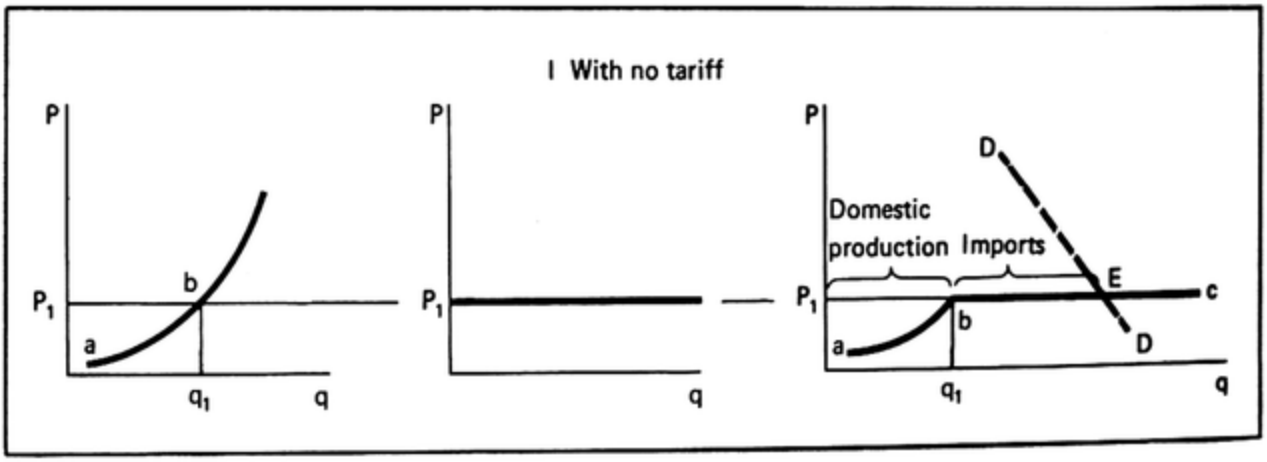
Domestic resource owners benefit from tariffs and quotas, particularly in the short run. In the long run, resources move

Tariffs and Quotas

This box will help you understand the detailed workings of tariffs and quotas. The first of its two diagrams illustrates a tariff.

Both panels of the first diagram show the determination of market price and quantity through the equality of supply and demand. The equilibrium point for the market as a whole is labeled with the letter *E*. The only tricky part is how the supply curves are derived.

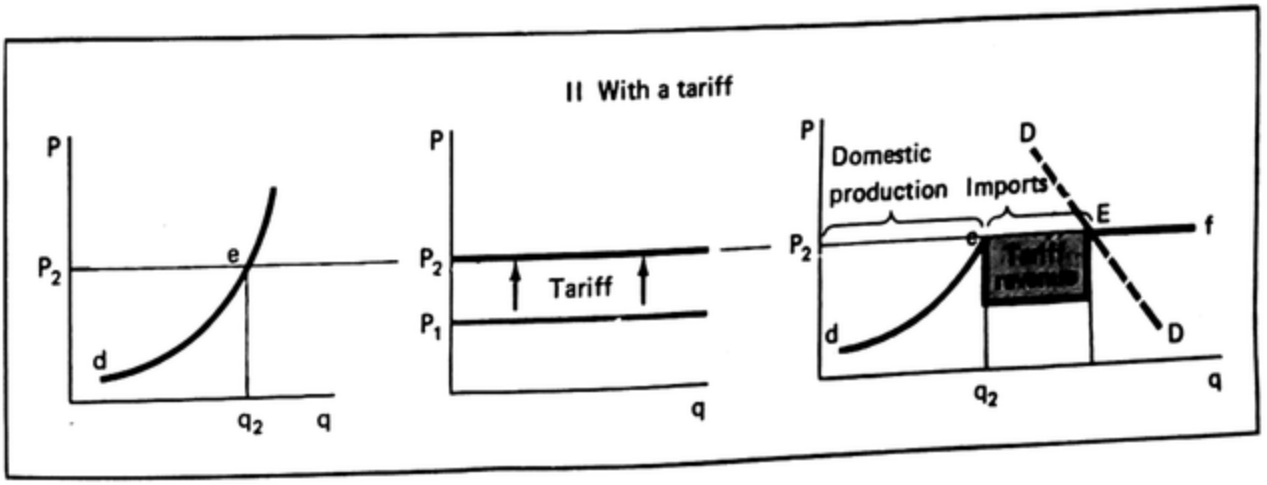
Look at Panel I of the first figure. There are two sources of supply, domestic and foreign. Domestic supply is only available at a cost that increases with

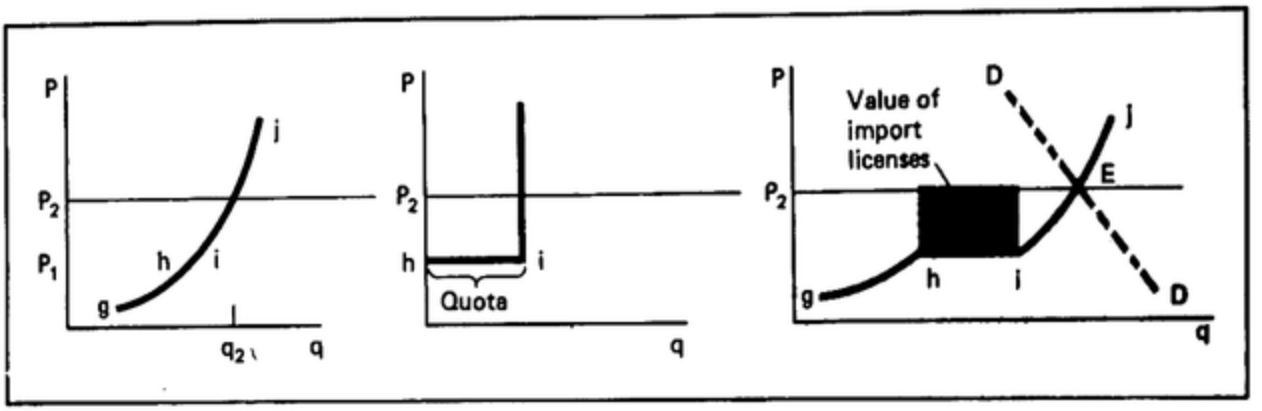


The effects of a tariff

Without a tariff (Panel I), the domestic market price is p_1. At that price, domestic suppliers produce q_1 and foreign sources supply the rest.

A tariff (Panel II) raises the cost of foreign goods from p_1 to p_2, and the domestic market price goes up by an equal amount. Domestic production rises to q_2, but imports drop off sharply. The effects of the tariff are to raise domestic price and output, but to reduce domestic consumption. The owners of domestic resources benefit, but domestic consumers suffer a loss.





The effects of a quota

An import quota that raises the price from p_1 to p_2 will have exactly the same effects on producers and consumers as a tariff equal to $p_2 - p_1$. The people who have import licenses will buy at the world price of p_1 and sell at p_2. They will reap a gain equal to the amount that would be collected by the tariff.

output, so that the domestic supply curve slopes up. Foreign supply is assumed to be infinitely elastic at the world price p_1. Up to a domestic output of q_1, domestic suppliers are the cheapest source. But at any higher output, they lose their comparative advantage because of increasing-cost, so that the constant-cost foreign supply is cheaper. The market supply curve is *abc*. The segment *ab* comes from the domestic supply over the range of its comparative advantage. The segment *bc* comes from the world supply curve. In equilibrium, the domestic industry supplies q_1, and the remaining demand is satisfied from the world market.

Panel II of the first figure is much like Panel I. A tariff raises the cost of foreign supply to p_2. Foreign suppliers continue to receive p_1, however, so that the tariff collected on each unit imported is equal to $p_2 - p_1$. Up to an output q_2, domestic supply is cheaper than foreign supply at the higher price, p_2. The market supply curve *def* consists of a domestic segment, *de, and a foreign segment, ef.* The difference, $q_2 - q_1$, consists of domestic supply produced at a comparative disadvantage. The total tariff revenue consists of the cross-hatched area equal to the volume of imports times the tariff collected on each unit imported.

The second figure illustrates the effect of a quota set equal to the amount of imports under the first figure's tariff. Up to an output q_1, domestic producers have a comparative advantage. Thereafter, the advantage lies with foreign producers. However, imports are limited to *hi* by the quota, so that any demand beyond q_1 plus *hi* has to be supplied inefficiently. The market supply curve consists of a portion, *gh*, along which domestic suppliers are efficient, a portion, *hi*, equal to the amount of the quota, and a portion, *ij*, along which domestic producers supply inefficiently. Since the lucky recipients of import licenses get to buy at the world price p_1 and sell at the domestic market price p_2, they collect a return on their import licenses equal to the first figure's tariff revenue.

into the protected industries, and the benefits may be competed away. But in increasing-cost industries, the owners of the fixed resources responsible for increasing costs will benefit permanently. The consumers of a good protected by a tariff or quota give up more of other goods in exchange for it, however, than they would have to give up without protection. And since the good is being produced by relatively inefficient domestic suppliers rather than by relatively efficient foreign suppliers, there is a *net loss* due to inefficiency. It is not simply a redistribution. This net loss is a compelling argument against protection.

The case for protection

One of the few sound cases for protection is the ***infant industry*** argument. It applies to a country that is in the early stages of industrialization. Many industries have a *minimum efficient scale* at the level both of the individual enterprise and of the industry as a whole. If the industry and its firms are smaller than efficient technology requires, the resources employed in them would be better employed elsewhere. Efficient foreigners could supply the product more cheaply. But if the domestic industry could only *become* large enough, it could compete on, or even take over, the world market. Although tariffs protect an industry that may be inefficient in its infancy, those same tariffs permit it to grow and develop into an efficient industry. A policy that looks misguided in the short run proves wise in the long run. The mature industry becomes an exporter, and the tariff is eliminated.

A country that protects its undeveloped industries is taking a current loss by giving up benefits from trade. It is betting on a future gain, in the form of a developed industry that is relatively more efficient than the industries that might have prospered under free trade. Arguably, the United States might still be no more than a supplier of timber, wheat, and cotton to industrialized Europe had it not been for the high tariffs of the 19th century. These tariffs enabled us to become one of the dominant industrial countries of the world. Doubtless we prospered as a consequence. But it does not follow that we prosper now from protecting mature industries that are no longer competitive.

If there is a net loss from the imposition of tariffs and quotas in mature industries, why do so many countries have them? As you can probably guess, the question is naïve. Protectionist legislation exists because some citizens can organize politically to support it. Tariffs and quotas bring short-run benefits to everyone who makes a living in a protected industry. They bring lasting benefits to those whose gains cannot be competed away, like owners of scarce farmland or other monopolies. These beneficiaries of protection find it easy to form political alliances to support their common interest. In the early 1980s, seven out of the ten cities in the United States with the highest unemployment rates were in Michigan. The United Auto Workers, some of the auto producers, their suppliers, and Michigan politicians of both the major parties had little trouble smoothing over their differences and lobbying for a quota on imports of Japanese cars.

Such political alliances are usually put together by groups that want protection for their own industry. Sometimes, however, an alliance is formed to fight *against* protection for other industries. When the British manufacturers of Ricardo's day banded together to form the Anti Corn Law League, they were fighting tariffs on grain. Their interest lay not in cutting the price of their breakfast cereal, but in cheapening the bread that formed the bulk of their workers' diets. This was

not philanthropy. Cheap bread meant low wages relative to the price of manufactures and high profits for the manufacturers.

When groups lobby for protection from foreign competition, they hardly ever come out and say, "We need protection to raise our incomes at the expense of those of our fellow citizens." Instead, they usually appeal to the patriotism of those very citizens at whose expense they hope to gain. The most common argument for protection is the claim that we need tariffs to shield American workers from the competition of "cheap foreign labor."

The claim goes something like this: Wages in most countries are lower than they are in the United States. How can high-wage American workers (and their employers, of course) compete with cheap Korean, Taiwanese, and Mexican labor?

This argument is wholly fallacious. Low-paid foreign labor does not produce cheap goods across the board. It is low paid largely because it is unproductive: poorly equipped and trained, and in poor health. Yet, compared with more productive American labor, it is at a *lesser absolute disadvantage* in some kinds of production than it is in others. This forms the *basis for a comparative advantage,* even though the poorer economies have an absolute disadvantage in all kinds of production. This comparative advantage leads to gains from trade. Real wages are raised in rich and poor countries alike, although the countries with unproductive labor continue to have lower wages than those with productive labor.

Does it follow that the auto workers of Detroit, Flint, Saginaw, and the other Michigan cities should have welcomed Japanese competition in the early 1980s? Of course not. They were greatly harmed, losing their jobs in the short run and taking relative wage cuts in the longer run. Japanese, German, and Italian competition broke down the monopolistic position of the American auto industry. Some of the industry's monopoly profits had been shared by the United Auto Workers. When the effects of foreign competition, high gasoline prices, and general economic stagnation hit all at once, the auto industry was devastated. Chrysler's workers soon accepted a pay cut to help the company stay afloat. The other companies' workers had to do the same a few years later. The economies of Michigan and other auto-producing states were in serious trouble.

Should the industry have been protected by either an import quota or a tariff? The answer depends largely on whether the industry would be able to retool to meet the competition. Industry advocates argued that the American firms were locked into a lot of capacity that had been designed for the production of larger cars. If they were given adequate time to rebuild their capacity and redesign their cars, they could meet the competition. Protection was needed in the short run to enable the industry to survive and become competitive again in the long run. This was not entirely a bad argument.

The situation of the auto industry in the early 1980s was complicated by several unique factors. First, a substantial part of the industry's problem stemmed from the oil price increase of the 1970s. The American auto companies were slow to shift to the production of fuel-efficient cars, a field that they had traditionally left to the Europeans and Japanese. Part of the problem, therefore, was genuinely transitional. There was no intrinsic, resource-related reason why the American companies could not compete in the small-car market. Second, wages in the American auto industry were out of line with those in much of the rest of American industry. Thus, part of the industry's cost disadvantage stemmed from relatively high wages rather than from productivity that was below that in other domestic industries. Third, the Jap-

anese cost advantage was partly the result of a pattern of export subsidy rather than lower comparative costs. Fourth, the most likely form that a reduction in the size of the American industry would take would be the bankruptcy of the Chrysler Corporation. This would seriously restrict competition in the (reduced) domestic market for large cars. Fifth, the industry was concentrated in a few geographical areas in which it was the dominant employer. Further decline in the industry would cause great hardship in these areas. Someone would have to bear the cost of relocating the affected workers to other regions of the country or equipping them to produce something else. These complexities had to be weighed along with arguments based on the theory of comparative advantage.

Very few public policy issues are simple. The argument on behalf of specialization according to one's comparative advantage is unassailable when it is confined to comparisons of alternatives at one moment of time. But it does not provide a conclusive guide to what to do in a changing historical situation in which there are costs of change and genuine uncertainties about where comparative advantage lies in the long run.

Export restrictions

The commonest restraints on trade are imposed by a country that expects to benefit its domestic industries by restricting import competition. Tariffs and quotas are restraints of this kind. Sometimes, however, countries deliberately restrict their exports by imposing taxes, quotas, or outright prohibitions on the export of particular goods. Since such a policy seems detrimental to the domestic export sector, it may strike you as perverse.

In fact, it sometimes makes perfectly good sense. In the early days of the Industrial Revolution, most European countries prohibited the export of new and specialized machinery that was used to produce other commodities for export. In effect, this preservation of trade secrets formed the basis for a national monopoly over the production of the goods using the machinery as inputs. This benefited the machinery-using industries by forestalling competition from foreign producers. Of course, it harmed would-be exporters of the machinery in question. To the extent that they were the same firms that used the machinery (as was frequently the case), the costs and benefits affected the same people. If a long-term monopoly was more beneficial to them than the short-term gains from selling machinery, they benefited from the restrictions.

In modern times, the commonest form of prohibition on exports is selective, not general. The selective restriction is used as a weapon of foreign policy rather than of domestic economic policy. The United States government must approve private sales of a wide range of strategic materials, weaponry, and nuclear fuels and facilities. It uses this power to reward the actions of some foreign governments and punish those of others. Sometimes the definition of strategic material is stretched a long way. In 1980, the U.S. government refused to approve grain sales to the Soviet Union because of its intervention in the civil war in Afghanistan. It also prevailed on the United States Olympic Committee not to "export" a team to the Moscow Olympics.

The most striking export restriction of recent times has been the ***OPEC oil cartel.*** A cartel is an association of producers formed to restrict the supply of a product to raise its price. The OPEC cartel sought to change the ***terms of trade***—to raise the prices of exports relative to the prices of imports, favoring the oil-exporting countries. An improvement in the terms of trade enables a country to raise its domes-

tic consumption and investment by getting more real imports for its exports. Since world demand for oil is very inelastic with respect to price, the cartel was spectacularly successful during the 1970s. By restricting their output, the OPEC countries succeeded in dramatically raising their potential to consume.

This is illustrated in Figure 5. Before the formation of the cartel, the typical member was producing on its domestic production possibility curve but consuming outside it because of moderate gains from trade. After the formation of the cartel, oil supply was restricted. The typical OPEC member was producing less oil and more other things. This was inefficient, but the change in the terms of trade that resulted from the oil supply restriction greatly improved the consumption possibilities of OPEC countries themselves. They had a common interest in maintaining their cartel to preserve the advantageous terms of trade that it made possible.

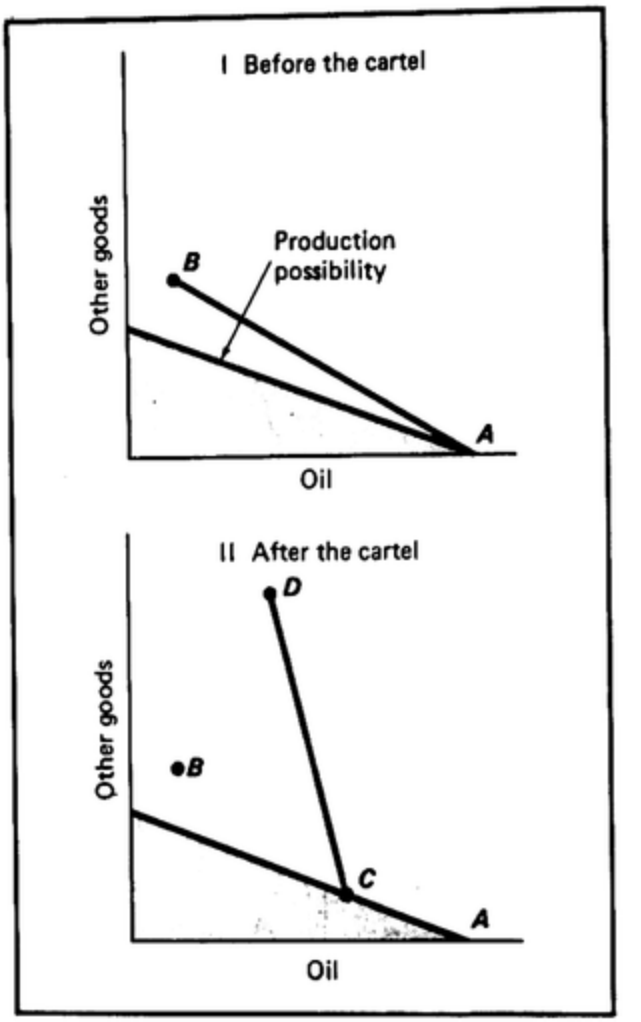


Figure 5 The effects of OPEC on its member countries

Before the cartel (Panel I), the typical member country was producing at Point *A* on its production possibility curve, specializing heavily in oil production. Because of moderate gains from trade, it could consume at *B*, outside its production possibility curve.

After the cartel (Panel II), the typical member country was producing less oil, at a point such as *C*. However, because of the enormous change in the terms of trade, it could consume at *D*,—that is, it consumed more of everything than it did at *B*, including more oil. Even though it was not exploiting its comparative advantage to the fullest, the rise in oil prices made possible by the reduced supply more than compensated the OPEC member for the inefficiency in production.

Foreign trade and the U.S. economy

Foreign trade and international finance have occupied a lot of front-page space in the newspapers over the past decade. To understand why, it will help you to know both the theories of trade and protection, which were the subjects of the first two sections of this chapter, and the theories of the balance of payments and exchange rates, which are treated in the next chapter. But without some knowledge of the facts of the past and how they relate to the theory, you will have trouble analyzing the future. Both this chapter and the next are designed to give you some of the factual background you need to apply economic analysis. The first part of this section surveys the history of total U.S. trade since the end of World War II. The second analyzes the composition of trade at the beginning of the 1980s.

The growing importance of trade

By the end of the 1970s, the dollar volume of trade, whether measured by exports or imports, was about two and a half times as large relative to GNP as it was in the early

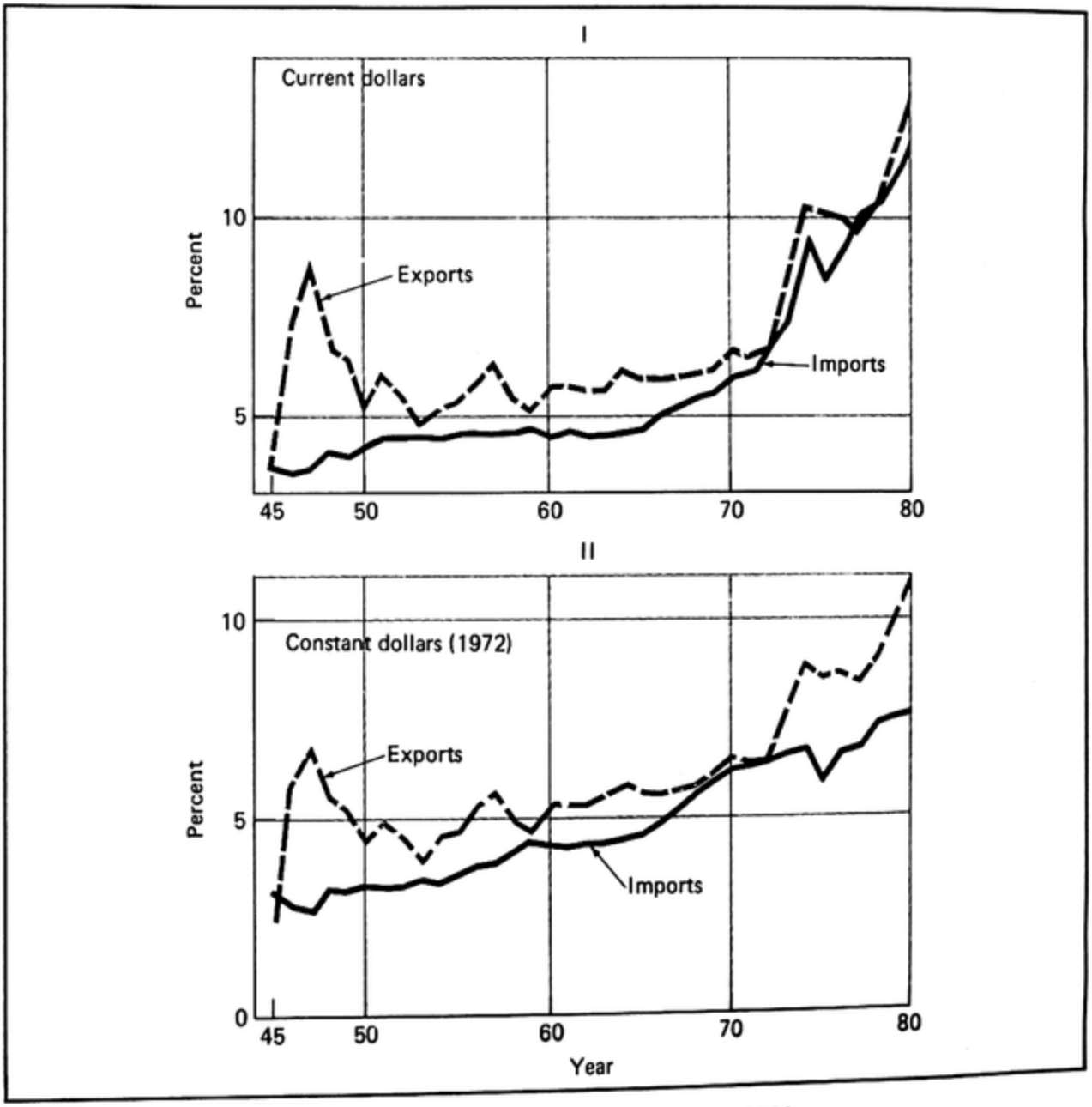


Figure 6 U.S. exports and imports as a percentage of GNP 1945–1980

In both current and constant dollars, U.S. trade has fluctuated widely relative to GNP in the 35 years since World War II.

Source: *Economic Report of the President.*

1950s. This is very large compared with most other changes in the structure of the economy. Much of it took place during the economically turbulent 1970s, but the dramatic rise in trade during these years reflected, in part, trends that originated much earlier. To set them in perspective, it helps to review trade developments since World War II. Look at Figure 6, focusing on the long-run trends rather than the year-to-year swings, which largely reflect business cycle movements.

Panel I shows exports and imports as a percentage of GNP, with all figures expressed in current dollars. Note the big bulge in exports right after the war ended

in 1945. Exports then dropped off until the mid-1950s, when they began a steady climb that accelerated after 1972. The overall picture for imports is broadly similar, except that it lacks the postwar bulge.

Panel II of Figure 6, in which both trade and GNP figures are expressed in constant (1972) dollars, tells the same story as Panel I up to 1973. Thereafter, there is a noticeable divergence. Although dollar imports rose rapidly relative to GNP, keeping pace with exports, real imports did not keep pace with real exports.

It is convenient for analyzing these trends to break the 35-year period down into subperiods. The first extends from 1946 into the early 1950s. Before World War II, England, the countries of Western Europe, and, to a lesser extent, Japan had been dominant industrial nations. They were far more specialized in manufacturing than the rest of the world. Their coal deposits, large urban labor forces, and highly developed capital and scientific facilities gave them a comparative advantage in heavy industry. Many of these countries, particularly France, also had highly productive agricultural sectors.

The destruction wrought by the war changed this. European and Japanese industry lay in ruins. The fighting had also interfered with the regular cycles of agriculture. With no manufactures to trade for food, Europe and Japan faced famine unless they could get help.

For both strategic and humanitarian reasons, the United States responded to the crisis with massive food shipments and assistance in rebuilding war-torn industry. It could do so because its own economy had developed very rapidly during the war and had been untouched by the fighting. As the immediate postwar economic crisis gave way to the Cold War, emergency relief assistance gave way to military assistance. From 1946 through 1951, the United States exported nearly $40 billion worth of goods more than it imported. Three fourths of these exports were paid for by U.S. government grants and loans to our former allies and enemies. All told, about 2 percent of U.S. GNP was transferred abroad in this way. This amounted to about 30 percent of our exports. Another 15 percent or so was financed by transfers of gold and other assets from Europe to the United States, so that only about half of our exports were paid for by imports.

Normally, what a country exports is roughly balanced in money terms by what it gets back in imports. Since the imports it gets are more valuable than those it could produce at home with the resources of its export industries, it gains from trade. This is established by the theory of comparative advantage. When it chooses, in effect, to give away its exports, it forgoes these gains. The United States government deliberately chose to do this after the war to speed the economic recovery and rearmament of the main industrial centers of the noncommunist world. The alternative would have been economic collapse throughout much of the world, including, of course, America's own export industries. Although the wisdom of rearmament was perhaps debatable, the decision to rebuild the economies of Europe and Japan produced many lasting benefits for the United States at a comparatively small short-run cost.

Next, look at the early 1950s through the mid-1960s. As you can see from Panel II of Figure 6, this was a period of rapid expansion in U.S. trade, which grew faster than U.S. GNP. The main factor responsible for this was the high rate of economic development in Europe and Japan. This provided expanding demand for U.S. exports and vast improvement in the variety and quality of manufactured goods that Americans could import from these countries. The result was an enormous flowering of trade among the industrialized

countries, greater specialization, and increased gains from trade. These gains were the payoff for the U.S. postwar aid program.

The real growth in trade continued to outpace GNP in the late 1960s and early 1970s, as you can see from Panel II of Figure 6. Two major factors contributed to this. The first was a substantial all-around reduction in tariffs negotiated in the mid 1960s, which encouraged further world trade and specialization. A second element was the spread of *multinational corporations*, enormous enterprises with branches in several countries. These companies' divisions are often closely tied to one another in related stages of production. Whenever they expand abroad in ways that either use American inputs or feed inputs into their American operations, our trade expands. For example, about 5 percent of U.S. commodity imports in 1979 were motor vehicles and parts from Canada. These were almost entirely the products of Canadian subsidiaries of GM, Ford, and Chrysler. A similar volume of trade in vehicles and parts crossed the border in the other direction. If the U.S. firms had decided to expand in Michigan and Ohio rather than in Ontario, or if the United States had captured Ontario in the War of 1812, much of this would have been internal trade.

There is an obvious difference, however, between the growth pattern of trade in the 1950s and early 1960s on the one hand, and in the late 1960s and early 1970s on the other. In the earlier period, export receipts ran consistently ahead of imports. During the later period, the gap between exports and imports gradually narrowed, largely because of the U.S. inflation that grew out of the Vietnam War. Before March 1973, the *exchange rates* among currencies of different countries were fixed by international agreement rather than determined by supply and demand. As U.S. prices rose relative to those of its international competitors, foreign imports became cheaper for Americans and U.S. exports more expensive for foreigners. This had nothing to do with changes in comparative advantage, which is rooted in real productivity differentials. It was purely a monetary phenomenon, reflecting overall price changes between countries rather than relative price changes between goods.

An ***export surplus*** (excess of exports over imports) is not in itself particularly desirable for the country that is running it. That country is giving up more in exports than it gets back in imports, which is not a very good deal. The reason for the surplus in the 1950s and 1960s was the flow of American investment into foreign assets. Much of this was investment by U.S. multinational corporations in their foreign subsidiaries. The dollars that Americans invested abroad enabled foreigners to buy more goods from us than we bought from them. The excess of exports over imports was compensated for by a growth in American-owned assets abroad. In a sense, the export surplus was a form of saving. Some forms of national income accounting refer to an export surplus as ***net foreign investment.***

In the late 1960s and early 1970s, when the export surplus narrowed, there was no corresponding reduction in investment by American firms in their foreign subsidiaries. Investment dollars were not offset by trade earnings, and there was an international financial crisis, which you will read about in the next chapter. One result was a major change in the international monetary arrangements for the settlement of trade and investment accounts. Exchange rates were set free to respond to supply and demand. This change had a big effect on international trade in the late 1970s.

Now turn back to Figure 6 and look at the late 1970s in both panels. In Panel I,

you will see a fairly close correspondence of exports and imports in current-dollar terms, except during the 1975 recession. In Panel II, you will see a widening export surplus in real terms. What does this mean?

In money terms, import growth nearly kept pace with export growth. In real terms, it did not. The only possible explanation is that import prices grew more rapidly than export prices. Between 1972 and 1980, prices of exports increased 110 percent; prices of imports increased 190 percent. This was an enormous swing in the terms of trade against the United States.

Two factors were at work. The first and most obvious was the formation of OPEC and the enormous rise in the price of oil. Since OPEC prices its oil in terms of the U.S. dollar, this was a direct jolt to the price of imports. A secondary factor was the drop in the value of the dollar. Both the increase in the cost of our oil imports and a continued increase in American investment abroad meant an excess supply of dollars on the world currency market. After the freeing of exchange rates in 1973, an excess supply of dollars caused a drop in the value of the dollar and a rise in the dollar price of imports other than oil. This restricted the growth in real imports. The other consequence of the depreciation of the dollar was a cheapening of American goods on the world market and a more rapid growth in real exports. A real export surplus opened up, even though the money values of exports and imports remained close to each other.

This was costly. An export surplus is a real cost to the domestic economy. The surplus that developed in the late 1970s resulted from a bad turn in the terms of trade, and thus it caused a loss of real income. Unlike the export surplus of the period immediately following World War II, it was not a deliberate decision to invest in the future of Europe, but a forced investment in the future of the OPEC nations—forced in part by OPEC, in part by our own inability to curtail the use of oil. The 3½ percent of 1980's real GNP that was shipped abroad in this export surplus was larger than the entire amount devoted to homebuilding in that year.

The composition of trade

Now that you know something about the volume of U.S. trade, you may be curious about its composition. What goods do we export and import? Who are our major customers and suppliers? These are not easy questions to answer, because the pattern of U.S. trade is very complex. This is an enormous, diversified country by comparison with nearly all of the other nations of the world. Carl Sandburg sang of this diversity in his poem "Chicago":

> Hog Butcher for the World,
> Tool maker, Stacker of Wheat.
> Player with Railroads and the
> Nation's Freight Handler.

Statistics are never as evocative as poetry, but they tell us that our country is both agricultural and industrial, that it exports complex aircraft and imports equally complex aircraft, that it is the world's second largest manufacturer of automobiles, and yet also its largest importer of automobiles.

To get some picture of a very complicated situation, study Table 2. It breaks down U.S. commodity trade in 1979 into 16 commodity groups and 5 regions. In discussing the content of this table, it will be convenient to lump the commodity categories into four groups of four categories each. The first four (food through fuels) will be called *primary products*, the second *semimanufacturers*, the third *machinery*, and the fourth *consumer goods*. These labels are not entirely accurate, but they make the discussion easier to follow.

Table 2 ***U.S. commodity trade by area and commodity, 1979 (in billions of dollars)***

Commodity	Area: Canada Ex	Canada Im	Japan Ex	Japan Im	Western Europe Ex	Western Europe Im	OPEC Ex	OPEC Im	Other Third World Ex	Other Third World Im	Total* Ex	Total* Im
Primary products												
Foods, beverages, tobacco	1.5	1.7	5.1	0.2	10.2	3.5	2.4	0.9	80.1	10.1	32.8	18.5
Raw materials	0.9	4.5	2.7	0.0	2.5	0.4	0.2	0.5	2.3	1.0	9.6	6.7
Ores and minerals	0.6	1.4	0.9	0.0	1.9	0.2	0.1	0.1	0.8	1.3	4.4	3.3
Fuels	1.5	5.5	1.2	0.1	1.8	3.3	0.1	40.1	0.9	10.7	5.6	60.0
Semimanufactures												
Nonferrous metals	0.5	1.7	0.4	0.3	1.1	1.4	0.1	0.1	0.4	2.1	2.5	6.4
Iron and steel	0.6	1.0	0.1	2.7	0.3	2.7	0.3	0.0	0.9	0.8	2.3	7.5
Chemicals	2.2	2.5	1.8	0.7	6.3	3.6	1.1	0.0	7.2	0.8	20.0	8.4
Other semifinished goods	1.2	3.4	0.5	0.6	1.5	2.2	0.5	0.3	1.6	2.4	5.6	9.8
Machinery												
Industrial machinery	4.4	1.8	0.8	1.2	4.0	4.8	3.0	0.0	6.1	0.5	19.9	9.2
Office equipment	1.1	0.6	0.7	2.0	4.9	1.0	0.5	0.0	2.5	3.7	10.1	7.4
Motor vehicles	8.7	8.0	0.2	9.5	1.4	5.7	1.6	—	2.4	0.3	14.7	23.6
Other equipment	5.0	2.9	1.9	3.0	8.8	5.0	3.4	—	8.4	2.7	29.0	13.7
Consumer goods												
Household appliances	0.5	0.2	0.2	3.4	0.8	0.9	0.3	—	0.8	2.4	2.5	7.0
Textiles	0.6	0.0	0.1	0.4	1.1	0.6	0.2	0.0	0.8	0.9	3.2	2.1
Clothing	0.1	0.0	0.1	0.3	0.5	0.0	0.0	—	0.4	4.7	0.9	5.7
Other consumer goods	1.1	0.8	0.4	1.1	1.8	4.6	0.5	0.0	1.1	5.0	5.2	12.2
Total*	31.2	37.6	17.3	26.2	49.2	41.4	14.4	42.0	45.2	50.3	170.6	204.9

Source: GATT, *International Trade, 1979–80.*

*Totals reflect areas and commodities not listed separately. Exports and imports of services (of which the U.S. is a substantial net exporter) are not included at all.

Look first at the last column, which shows export and import totals for the commodity categories. Note that the United States trades most heavily in machinery and primary products, somewhat less in semimanufactures, and relatively little in consumer goods. It is a substantial net exporter of food, chemicals, industrial machinery, and other equipment, and a substantial net importer of oil, metals, motor vehicles, and consumer goods.

Now look at the individual columns. U.S. trade with Canada, as you know, is heavily influenced by the internal trade of multinational automobile firms, which accounts for about a fourth of the trade between the two countries. The other trade between them consists mainly of exports of machinery from the United States and of primary products from Canada.

Trade between the United States and Japan is also relatively simple. There is a substantial net outflow of primary products from the United States, and of steel, machinery, automobiles, and home appliances from Japan. It may strike you as interesting that with Canada, the U.S. exchanges manufactures for primary products, but in its trade with Japan, it exchanges primary products for manufactures. Canada has a substantial comparative advantage in grain, timber, pulp, minerals, and natural gas. Japan is highly specialized in manufactures, since it has a large skilled labor force but lacks land and natural resources. The United States, with its great diversity, is somewhere in between. Its primary products industries export to Japan; its manufacturing industries export to Canada. In the spectrum of comparative advantage, it is located between these two trading partners.

Trade with Western Europe is more complicated. The United States is a heavy exporter of primary products—mostly food, tobacco, and industrial raw materials. However, it also imports a lot of primary products, especially wines, spirits, and petroleum products. There is a substantial trade in both directions in semimanufactures and machinery, with the United States a net exporter of chemicals and equipment, and Western Europe a net exporter of steel and motor vehicles. Finally, these two affluent industrial areas enrich each other's consumption patterns in various ways. Scandinavian furniture is particularly nice.

Trade between the United States and the OPEC countries follows a classic pattern of comparative advantage. They send us oil; we send them food and machinery. The pattern of trade with other Third World countries is less tidy, however, partly because these countries are such a diverse collection. Taiwan and the Republic of Korea are quite industrialized. We send them food, chemicals, and machinery, and get back equipment, appliances, clothing, and other goods. The countries of Africa, South Asia, and Central and South America are less industrialized. We send them chemicals and machinery and get back primary products.

This whole pattern of trade is complex, almost bewildering. It may seem to you that it is so far removed from the wheat-tulip world of Belgium and Holland that its organizational principles must be far removed from the law of comparative advantage. If you think so, you are wrong. We do not import airliners from Sri Lanka, nor do we ship coals to Newcastle. Most of the seeming contradictions of American trade can be resolved if you remember that this is both an agricultural and an industrial country. Nowhere is this paradox more clear than in California. Just south of San Francisco is "Silicon Valley" (the Santa Clara Valley, in fact)—the world capital of semiconductor technology, the scientific basis of the coming second indus-

trial revolution. Less than 100 miles away is Castroville, the "Artichoke Capital of the World."

Summary

This chapter covers the real, underlying basis for international trade—the patterns of productivity and demand that determine who will export to whom, and who will benefit the most from it. Its principal lessons are these:

1. Countries benefit from international specialization and exchange, just as individuals benefit from the division of labor.
2. Countries tend to specialize in those commodities in which they have a comparative advantage, that is, in whose production their resources are relatively productive. Even if one country's resources are more productive than another's across the board, it can benefit from specializing in those activities in which it has the biggest advantage. By specializing in those commodities in which it has the smallest disadvantage, the less productive country also gains from exchange.
3. Although the world as a whole benefits from specialization and trade, the gains are distributed unequally. Countries that have a comparative advantage in products that are in great demand benefit the most. Those with an advantage in products that are relatively plentiful benefit the least.
4. Tariffs, quotas, and other forms of trade restriction lead to an inefficient pattern of world production. However, they benefit particular industries, and sometimes their countries. It is not surprising, therefore, that trade restriction is quite common despite the arguments against it on efficiency grounds. A spectacularly successful trade restriction was the formation of the OPEC oil cartel in the 1970s, which reduced world oil supply and turned the terms of trade in favor of its members.
5. U.S. foreign trade has grown enormously in relative importance since World War II. However, in the past decade, the terms of trade have turned against the United States, and we benefit much less from trade than we did before the formation of OPEC.
6. The United States, with its great regional diversity, is both a major producer of primary products and a major industrial nation. In general, we trade manufactures to countries that specialize in primary products, and trade primary products to countries that specialize in manufacturing.

Key concepts

Law of comparative advantage
Absolute advantage
Gains from specialization and trade
Multilateral trade
Protectionism
Tariff
Quota
Infant industry
OPEC oil cartel
Terms of trade
Export surplus
Net foreign investment

Questions for review

1. a. Define: absolute advantage in trade and comparative advantage in trade.
   b. Explain carefully *why* trading patterns should be determined on the basis of comparative advantage rather than absolute advantage.
2. You are discussing the material on trade with a friend. Your friend is convinced that he has mastered the theory of trade. According to him, if one country can produce all goods more cheaply than a second country, no trade between the two countries should take place. There is, after all, no gains from trade for the country that can produce all goods at the lowest cost. Explain to your friend why he isn't as much of an expert on trade as he thinks he is.
3. Explain the impact on the net gains from trade of:
   a. increasing returns to scale
   b. decreasing returns to scale
   c. transport costs
4. A certain country exports coal across one border and imports it across another. Under what conditions would this be economically efficient?
5. In the 1980s, the UAW lobbied for trade restrictions on the importation of foreign cars. Discuss some of the costs and benefits of such action. Under what conditions would such trade restrictions be most economically efficient?
6. Suppose that you are a trade adviser in a less-developed country. Spokespeople from a certain industry have asked the government to place a tariff on imports of the good that they produce. Their argument in favor of the tariff is the *infant industry* argument. In a summary written for your Minister of Trade:
   a. Explain the infant industry argument.
   b. Point out some issues that should be researched.
7. In the post–World War II period, there were massive exports from the United States to the war-torn economies. Since most of these goods were given away, the usual gains from trade would seem to have been forfeited. Yet, it can be argued that this policy of export grants and loans was essential to the long-run economic strength of the United States. Explain.

35

International Finance

As you read and study this chapter, you will learn:

- how people and firms in one country make monetary payments to those in another
- what determines the international value of the dollar
- what the balance of payments is, and why it is sometimes a major economic problem
- how the world economy has adjusted payments imbalances under various systems of currency exchange

Most of us expect to get paid for what we produce, or at least to keep our products. This is such an ingrained feature of our society that it is hard to imagine how things could be otherwise. It may surprise you, then, to know that many societies share their produce according to quite different principles. Our idea of exchange is to trade one thing either for another thing or for money, which is the power over other things. Anthropologists call this "balanced reciprocity." But they have discovered other forms of reciprocity. One of the most interesting is the "generalized reciprocity" of the Kung Bushmen, who live in South Africa. When the hunters of this society kill animals, they give the meat away—not all of it, but most. They expect nothing in return and, indeed, would be insulted by thanks, not to mention pay. The people who get the meat subdivide it and, again, give it away. Visitors are welcome to a portion. No one expects anything immediate in return. It is all done according to a pattern of obligation, in which each tribal member must share with the others.

Reciprocity arises only because the others are similarly obliged, and the giver can expect to be supported in part from the gifts of others. In a way, it is like multilateral trade. A trades to B, who trades to C, who trades to A. But in trade, money changes hands. In generalized reciprocity, it is simply understood that membership in society entitles a person to receive and requires him or her to give. In spirit, it is much like what goes on in a family whose members care for one another.

International trade, of course, is based on balanced reciprocity. Sellers expect to be paid, and buyers expect to pay. In this respect, it is no different from domestic trade. But the existence of national currencies creates a complication in international exchange that is not there in domestic exchange. An American who buys a Mercedes pays for it with dollars. When the Mercedes dealer replenishes his inventory, he must somehow convert his dollars into marks, since the manufacturer, Daimler Benz, expects to be paid in German currency. Dollars are of little use to the German automaker, since it must pay its workers, suppliers, and stockholders in marks.

International finance is the study of international monetary transactions. It is closely tied to the study of international trade, since most international monetary transactions are made to settle trade debts. However, international capital movements are also central to monetary relations among countries. Individuals and firms from one country often want to invest in another. To do so, they must convert their own currency into that of the country in which they are investing. Although capital flows are not as large as trade flows, they are highly volatile. Much of the "action" on the international currency markets is related to sudden reversals in the flow of private capital from one country to another. Finally, governments also deal in international currencies, partly as buyers and sellers of commodities, partly as borrowers and lenders, and partly as interested parties in the stability of their own currencies.

This chapter is divided into two major sections. The first describes and analyzes the U.S. ***balance of payments,*** which is made up of the various demands for and supplies of dollars on the world currency markets. The second describes the institutions and workings of the ***world payments system,*** which is the set of international arrangements for bringing demands for and supplies of the world's currencies into balance with one another.

International currencies and payments

When Americans—individuals, firms, or the government—want to convert their dollars into foreign currency, they must work through international financial intermediaries. Their banks carry out the conversion for them. If the banks themselves do not hold balances in other banks abroad, they must sell dollars and buy foreign currency on the world ***currency exchanges,*** or *currency markets.* Thus, the Americans who "convert" their dollars into other currencies are indirectly selling dollars and buying pounds, yen, or marks. Similarly, when foreigners need dollars, they go through intermediaries to sell their own currencies and buy dollars. Of course, firms and wealthy people who regularly buy and sell in various countries often keep bank accounts in all of them, to avoid having to go through the currency markets for every international transaction. But whenever they want to transfer funds from one country to another, they must use the currency exchanges.

The currency exchanges are really major banks located in such places as New York, London, Zurich, and Tokyo. They quote buying and selling prices at which they will trade one currency for another. Thus, the banks "make a market" for each currency that is exchanged and establish market-clearing prices, or ***exchange rates,*** in response to supply and demand.

Currency markets

A currency market is a lot like any other competitive market, except that money is exchanged for other money rather than for goods and services. The demands for and supplies of currency are, therefore, derived entirely from their ability to buy other things, and not from any direct usefulness. This affects the economic logic that lies behind the supply and demand curves.

Suppose for the moment that there are only two currencies, the American dollar and the Japanese yen. What would the currency market look like? To picture it, focus on the dollars demanded and supplied and on the number of yen that it costs to buy a dollar. Figure 1 illustrates the dollar market by means of a demand-and-supply diagram. As it is drawn, the demand curve slopes down and the supply curve slopes up, although the supply curve may bend backward.

The downward slope of the demand curve derives from the downward slope of the Japanese demand curve for American goods. Given the domestic price levels in the two countries, as the dollar gets cheaper in terms of yen, American goods become cheaper for the Japanese. The Japanese demand more dollars to get more dollar goods. Similarly, the supply curve of the dollar slopes up, because Americans will supply more dollars in exchange for yen when the dollar is more valuable. Its greater value enables Americans to buy more Japanese goods.

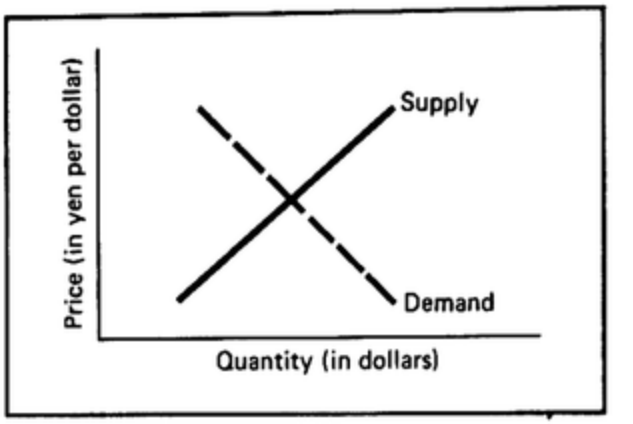


Figure 1 Demand for and supply of the dollar

The demand for the dollar goes up as its price (in terms of yen) goes down, because the Japanese will want more dollars when they are cheaper, to buy more American goods, which get cheaper as the dollar itself gets cheaper. The supply of the dollar goes up when its price goes up. Americans supply more dollars to get more yen, since when the dollar is more valuable, the yen is cheaper, and so are Japanese goods.

The supply curve might slope down instead of up because of the underlying American demand for Japanese goods. If the dollar gets more valuable relative to the yen, Americans will buy more of the cheaper Japanese goods. But they may spend fewer dollars to buy more goods, since a dollar will buy more yen. This will make the supply curve for dollars slope "backward."

What matters, however, are the relative slopes. In both parts of Figure 2, a price equal to P_o clears the market. At any higher price, there is excess supply, and price is depressed toward P_o. At any lower price, there is excess demand, and price is driven up toward P_o. Thus, demand and supply tend to establish an equilibrium exchange rate for the dollar in terms of the yen.

The world has many different currencies, exchange rates, and currency markets. The markets and exchange rates are tied to one another in a pattern by ***arbitrage,*** which is the process of "buying cheap and selling dear." Suppose that on the Zurich exchange $1.00 buys 5 Swiss

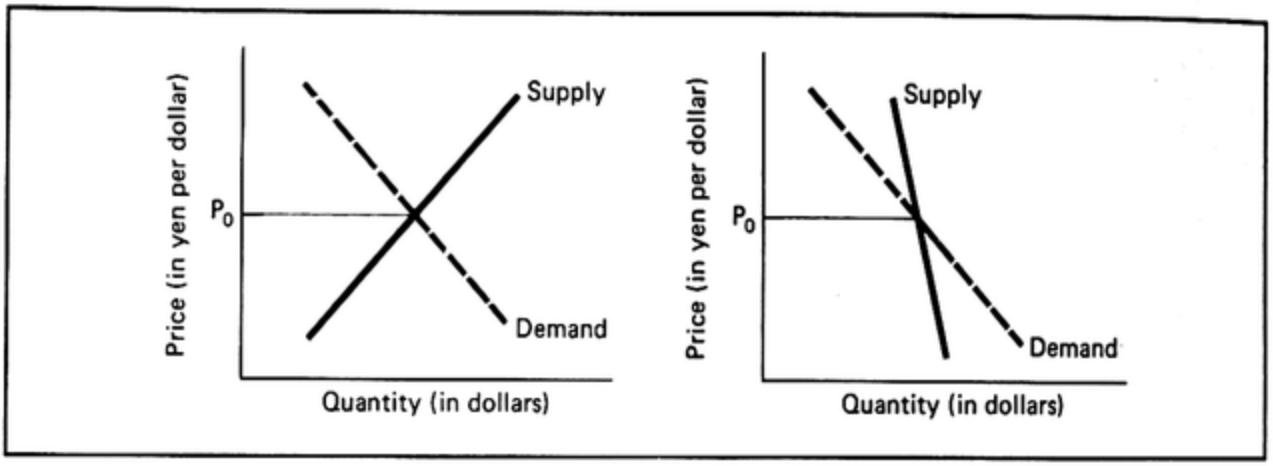


Figure 2 Stable markets for the dollar

In both of these diagrams, a price, P_0 clears the market for dollars. At a higher price, there is excess supply. At a lower price, there is excess demand. Thus, the markets tend to settle at a competitive equilibrium P_0.

francs, 5 Swiss francs buy 3 marks, and 3 marks buy $1.50. Arbitragers will simultaneously sell dollars for francs, sell francs for marks, and sell marks for dollars. They will continue to do this in a large way as long as it is obviously profitable—for example, until the mark falls in value, so that 3 marks purchase only $1.00. Then prices will be back in line, and it will no longer be possible to profit from a three-way transaction. Similarly, if the dollar is cheaper relative to the mark in London than it is in Hamburg, arbitragers will simultaneously buy dollars for marks in London and sell them in Hamburg. This will bring their prices back into line on the two exchanges. It is not ***speculation,*** not a matter of uncertainty, since arbitrage involves simultaneous buying and selling of a currency at two different prices. It is a sure thing. The mere possibility of arbitrage keeps exchange rates in line around the world, since enormous amounts of money can be mobilized in response to a certain gain.

The balance of payments

Trade in goods and services is the main source of demands for and supplies of international currencies. But another major reason for transferring purchasing power across national boundaries is financial investment. Private individuals, firms, and governments in one country want to acquire financial assets in another. To do this, they have to go through the currency exchanges.

Because of this additional dimension, it is a lot easier to understand the currency markets if we have a systematic method of accounting for all demands and supplies. This is, in fact, provided by *balance-of-payments accounting.* It gives us a map of the complex pattern of currency flows and is as important for understanding the international pattern of payments as national income accounting is for understanding how GNP is determined.

The U.S. balance of payments is the pattern of transactions in the U.S. dollar on the world currency markets. Figure 3 presents a picture of this pattern as it looked in 1979.

Demands for and supplies of the dollar are grouped into three categories: the current account, the capital account, and official settlements. The ***current account*** consists largely of imports and exports, plus small amounts of transfer payments (such as foreign aid, pensions paid to Americans living abroad, and remittances that immi-

grant workers send home to their families). The ***capital account*** consists of lending and investment across national boundaries, some of it undertaken by individuals, some by multinational corporations, and some by governments. ***Official settlements*** transactions are government interventions in the currency markets, aimed at stabilizing currency values.

Table 1 shows the same information in greater detail. As you can see, the current account is dominated by merchandise trade, or trade in goods, but trade in services and income payments are also important. Income payments are included in the current account because they are payments for the use of borrowed funds and invested capital. As such, the payment of either interest or capital income represents the import of a capital service.

Changes in the *amounts of outstanding loans and investments* make up the capital account. When U.S. capital flows abroad, this supplies dollars to the rest of the world. Similarly, when Volkswagen invests in the United States, it demands dollars.

The official settlements account includes official government acquisitions of gold, foreign currencies, and other international ***monetary reserve assets.*** When the U.S. government buys such assets, it supplies dollars to the rest of the world. When it sells them, it reduces the supply of dollars. When foreign governments purchase U.S. Treasury securities or make de-

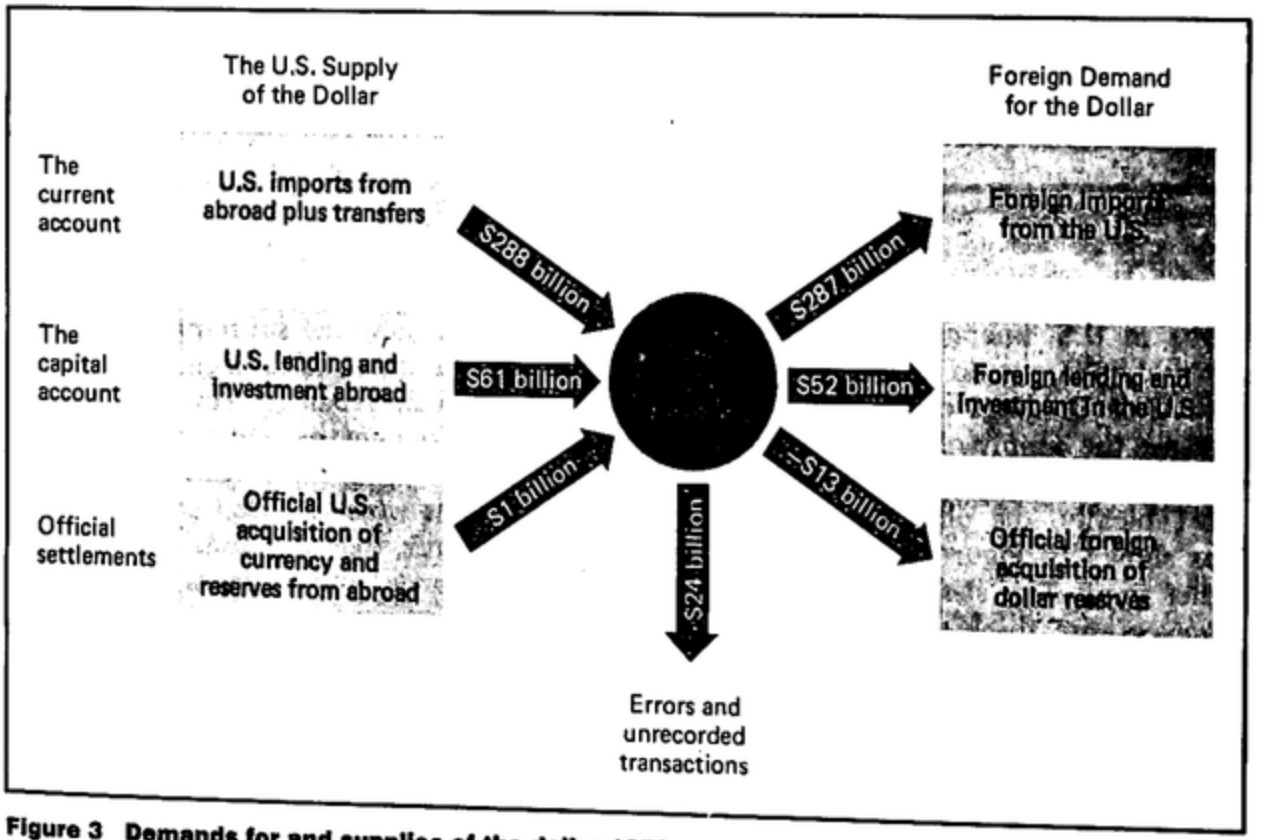


Figure 3 Demands for and supplies of the dollar 1979

Transactions in the dollar take three major forms: (1) imports and exports, (2) lending and investment, and (3) official reserve transactions. Any discrepancy in the supply-demand balance reflects measurement errors and unrecorded transactions.

Source: *Survey of Current Business.*

Table 1 ***U.S. International transactions, 1979 (in billions of dollars)***

Supply of the Dollar		Demand for the Dollar	
	The Current Account		
U.S. merchandise imports	211.5	U.S. merchandise exports	182.1
Defense expenditures abroad	8.5	Military sales	7.5
Travel, transportation, and other services	28.2	Travel, transportation, and other services	31.3
Income on foreign-owned assets in the U.S.	33.5	Income on U.S.-owned assets abroad	66.0
Government and private transfers	6.0		
	$287.6*		$286.8*
	The Capital Account		
U.S. lending and investment abroad			
government	3.8	Foreign private lending and	
private	56.9	investment in the U.S.	51.8
	$60.7		$51.8
	Official Settlements		
Official U.S. acquisition of gold, foreign currencies, and other reserve assets (net)	$1.1	Official foreign acquisition of dollar assets	−$13.2
	Errors and unrecorded transactions		$23.8
Totals	$349.3		$349.3

Source: *Economic Report of the President.*
*Totals do not add exactly due to rounding.

posits in U.S. financial institutions, they demand dollars to do so, but when they reduce their holdings of dollar assets, they increase the supply of dollars flowing into the world currency exchanges.

Typically, the official settlement figures shown in Figure 1 and Table 1 are rather small. This should not mislead you, however, into thinking that they are unimportant. Such transactions are the major means by which governments try to stabilize the markets on which dollars and other currencies are traded. When these markets are functioning normally, the amount of official intervention is small. Only when the currency markets are troubled are these transactions relatively large.

Finally, note the substantial entry for errors and unrecorded transactions, the "statistical discrepancy." This simply measures the amount by which the accounts fail to balance. If all international dollar transactions were recorded and measured without error, then demands and supplies would balance, and a separate balancing entry would be unnecessary.

Demands and supplies must be equal because a balance-of-payments account shows something that actually happened. Every dollar came from somewhere and went somewhere, whether or not the dollar market was in equilibrium. If it was not in equilibrium, someone's plans were frustrated, and his or her dollar acquisitions

were different from what was desired. Remember that national accounts must also balance, whether or not the multiplier process is in equilibrium. If it isn't, then actual and desired investment don't match. Similarly, if there isn't equilibrium in the balance of payments, there must be an unintended flow of dollars, or an official settlements transaction to take up the difference. The next step is to learn what the determinants of demands and supplies in the dollar market are. The rest of this section is devoted to the study of trade and capital flows.

The current account

Merchandise trade is the heart of the current account, its largest component, and also the major source of its instability. You can see this instability in Panel I of Figure 4, which shows the value of U.S. merchandise exports minus the value of imports from 1960 to 1980. By contrast, the balance of other current-account items was stable and close to zero until 1973, when a combination of ever-growing U.S. investment abroad and high profitability swelled a steady uptrend in net income from foreign investment.

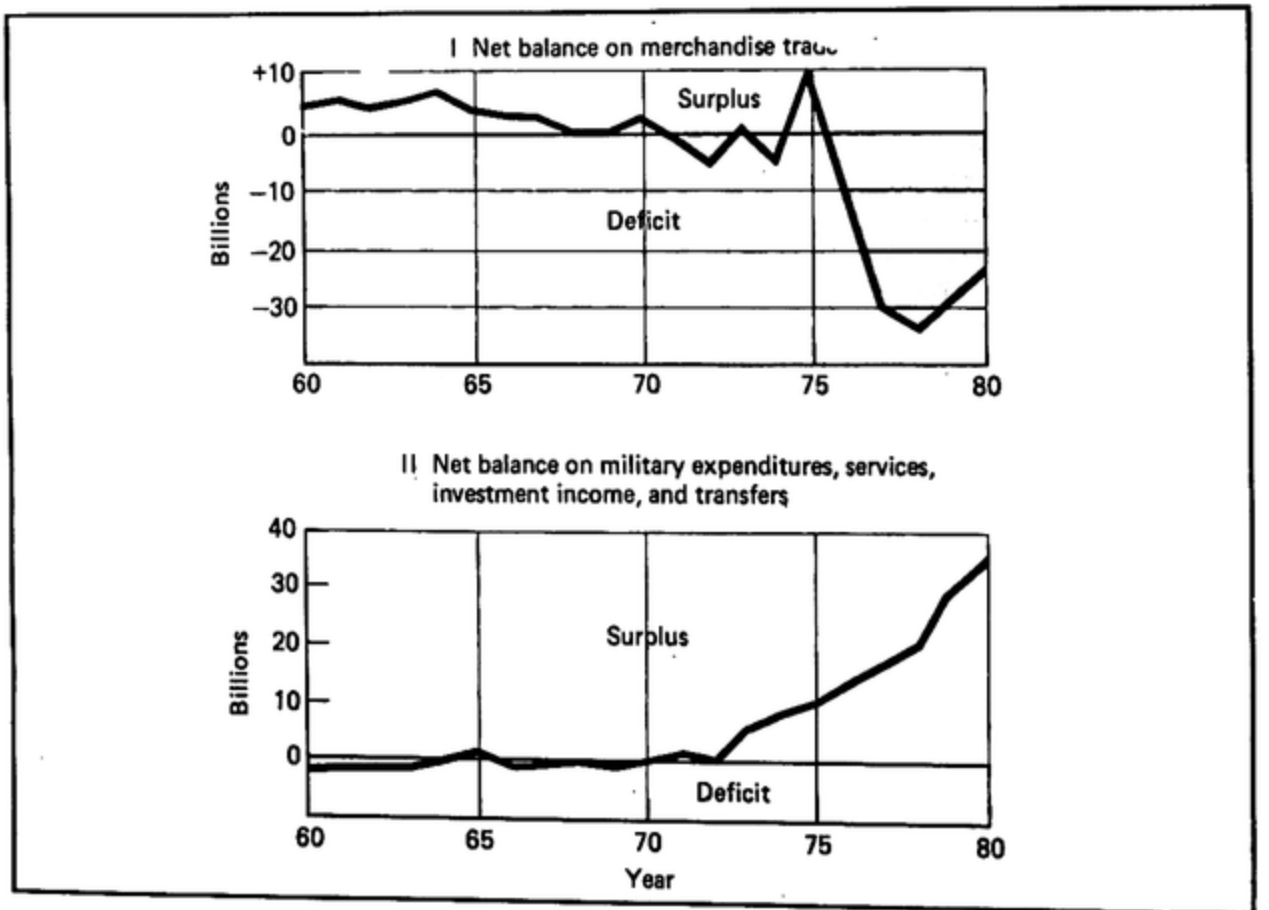


Figure 4 The U.S. current-account balance 1960–1980 (in billions)

Short-run fluctuations in the current account have been dominated by merchandise trade, but the balance on other current-account items swung sharply into surplus in the 1970s, mainly because of an uptrend in investment income.

Source: *Economic Report of the President.*

Since merchandise trade is so important and volatile, this section concentrates on it and ignores the other elements in the current account. It also ignores changes in comparative advantage and focuses, instead, on the effects of changes in demand. Year-to-year fluctuations in trade are dominated by changes in national incomes and price levels, not by changes in comparative advantage, which take place slowly.

To see how price and income changes affect patterns of trade, it will help you to look in greater detail at trade between two countries. This time they are the United States (U.S.) and the United Kingdom or Britain (U.K.). They trade with each other, but have different currencies. The U.S. currency is the dollar, and the U.K. currency the pound. Each country prices its goods in its own currency, so that trade between the countries requires that the currencies can be traded for one another on a currency exchange. When the U.S. imports from the U.K., dollars are supplied and pounds demanded. When the U.K. imports from the U.S., pounds are supplied and dollars demanded. The rate of exchange between the two currencies can be looked at from the perspective of either country. An American importer is concerned with the dollar price of the pound, with how many dollars must be paid for a pound. The U.K. importer is interested in the pound price of the dollar.

American exports equal British imports, whether measured in physical quantities, dollars, or pounds, since U.S. exports and U.K. imports are the same thing. This is tautological. But the condition for *balance* in the merchandise trade account is not tautological. One way to write it:

$$\begin{pmatrix}\text{Quantity of}\\ \text{U.S. exports}\end{pmatrix} \times \begin{pmatrix}\text{Dollar price}\\ \text{of U.S. exports}\end{pmatrix} = \begin{pmatrix}\text{Quantity of}\\ \text{U.S. imports}\end{pmatrix} \times \begin{pmatrix}\text{Dollar price}\\ \text{of U.S. imports}\end{pmatrix}$$

This says, in effect, that the demand for the dollar coming from our exports matches the supply of the dollar coming from our imports. A similar expression applies to Britain.

Suppose the two countries' merchandise accounts balance. What will happen if prices and incomes change in the two trading countries? The answer depends on the following economic factors:

The quantity of U.S. exports	depends on	1. British GNP 2. Prices of U.S. goods in the U.K. 3. Prices of British substitutes
The quantity of U.S. imports	depends on	1. U.S. GNP 2. Prices of U.K. goods in the U.S. 3. Prices of American substitutes
The dollar price of U.S. exports	depends on	1. The general level of U.S. prices 2. Special factors in the U.S. export industries
The dollar price of U.S. imports	depends on	1. The general level of U.K. prices 2. Special factors in the U.K. export industries 3. The dollar price of the pound

Apart from special factors (like the formation of a cartel in an export industry), what matters are GNPs and price levels in the two countries, and the rate of exchange between their currencies. Moreover, the important variables are the *relative* sizes of real GNP in the two countries and the *relative* price levels adjusted for exchange rates. If money GNPs in the two countries both change in a given proportion, but the two price levels change in the same proportion, the quantities of goods exchanged will not vary.

Changes in GNP

The effects of changes in GNP are straightforward. Suppose that America's real GNP drops and British GNP remains unchanged. What will happen to the relative demands for and supplies of the two currencies? American demand for British goods will drop; so will the supply of dollars that Americans wish to convert into pounds. But there will be no change in British demand for U.S. goods. The supply of pounds demanding dollars will also be unchanged. There will be an excess supply of pounds and an excess demand for dollars.

You can see the effects of a GNP drop in Panel I of Figure 4. The sharp peak in 1975 in the U.S. merchandise trade balance came from the steep U.S. recession in that year, which had a pronounced effect on merchandise imports. As the GNP fell, our trade account swung into surplus, producing an excess demand for the dollar from this source.

What if U.S. and British GNP were both to change in the same proportion? Would both countries' imports rise to keep the current accounts in balance? This would depend on the countries' *income elasticities of demand*—the percentage change in demand resulting from a 1 percent change in income. In this context, the relevant elasticity is that of imports relative to GNP.

Suppose that the demand elasticity of U.S. imports with respect to its GNP is 2. Then a 10 percent increase in GNP results in a 20 percent increase in imports and a 20 percent increase in the supply of the dollar. If the elasticity of British import demand is 1 rather than 2, a 10 percent rise in British GNP produces only a 10 percent increase in import demand and a 10 percent increase in the demand for the dollar. If both countries' GNP levels grow at the same rate, the United States will find its currency in excess supply. If both countries have *equal* demand elasticities, the one whose GNP is growing faster will increase the supply of its currency relative to the demand for it on the part of its slower-growing trading partner.

The history of Japan during the 1960s and early 1970s illustrates these principles well. From 1960 to 1974, Japanese industrial production increased at an annual rate of 7 percent, just about twice the rate of increase in the output of most of the other major industrial countries. Yet, the Japanese current account remained in balance. The main reason for this was the exceptionally high world income elasticity of demand for high-quality manufactured goods. Even though income in the rest of the world was growing more slowly than it was in Japan, the demand for Japanese exports kept pace with Japan's rapidly growing imports of food and raw materials.

Changes in relative prices

The effects of price changes on the current account might seem complicated at first because of the possibilities to be considered. American prices can change, British prices can change, or the exchange rate can change. The fact that it is relative prices that matter greatly simplifies things. Note that each of the three events listed below raises the price of American goods relative to British goods by 10 percent in both countries:

1. A 10 percent increase in all U.S. prices.
2. A 10 percent drop in all U.K. prices.
3. A 10 percent fall in the dollar price of the pound, which is also a 10 percent rise in the pound price of the dollar.

Because all three have equivalent effects on relative prices, they also have equivalent effects on the demands for and supplies of the two currencies.

To see how these demands and supplies will change, consider a particular example. Suppose that all U.S. prices go up by 10 percent and both U.K. prices and the exchange rate remain constant. What will happen to the supply of and demand for dollars on the world currency market?

It might seem at first glance that a rise in U.S. prices would necessarily increase the dollars that Americans supply in exchange for pounds and reduce the dollars that the British demand. Thus, it would create an excess supply of dollars on the world market and a corresponding excess demand for the pound. In fact, however, a rise in the relative prices of American goods could create an excess demand for dollars and an excess supply of pounds. Conceivably, the balance of supply and demand in the currency markets could even be unaffected by the change in relative prices. The key to what happens lies in the *price elasticities of import demand* in the two countries. The American price elasticity is the percentage change in American demand for British goods resulting from a 1 percent change in the relative price of British goods. The British demand elasticity is similarly defined. These two elasticities determine what will happen if U.S. prices rise by 10 percent, making British goods relatively cheaper in both countries.

First consider U.S. imports. A 10 percent rise in U.S. prices is the same as a 10 percent drop in the relative price of British goods. Since British goods are now cheaper, U.S. imports will rise by an amount determined by the U.S. import demand elasticity. Suppose it is 0.7. Then a 10 percent drop in the relative prices of British goods will result in a 7 percent rise in the *quantity* of U.S. imports. At the prevailing exchange rate, the 7 percent increase in U.S. imports results in a 7 percent increase in the dollars supplied to pay for them.

Now look at what happens in the U.K. Since a rise in U.S. prices, other things being equal, causes American imports to be relatively more expensive in British markets, the quantity of goods imported into Britain will fall. If the British import demand elasticity is, say, 0.3, then the *quantity* of British imports will drop by 3 percent. But since the dollar price of imports has risen 10 percent, the dollars needed to pay for the smaller quantity will actually rise by 7 percent (10 percent higher prices less 3 percent lower quantity). Thus, in this example, the British demand for dollars rises in exactly the same proportion as the U.S. supply. The U.S. is spending 7 percent more to import a larger quantity of British goods. The British are paying 7 percent more to import a smaller quantity of the more expensive American goods. Imports change in both countries, but the balance of the current account is not changed.

These figures were deliberately chosen to make a point. It is entirely possible that a change in relative prices will leave the balance of the currency market unaffected. All that is required is that the two elasticities sum to 1 ($0.3 + 0.7 = 1$). But if this is not the case, then a change in relative prices will cause an excess demand for or supply of the dollar. For example, if either of the demand elasticities had been smaller, there would have been an excess *demand* for dollars. Suppose that the U.S. demand elasticity had been 0.5 rather than 0.7. Then the supply of dollars would have risen by 5 percent and the demand by 7 percent. There would have been excess demand. If the British demand elasticity had been 0.2 rather than 0.3, demand for dollars would have risen by 8 percent rather than 7, and again, there would have been excess demand.

By the same token, if the sum of the demand elasticities had been greater than

1 in magnitude, a rise in the American price level would have created an excess *supply* of dollars. You should verify this by running through the previous example with an American elasticity higher than 0.7 or a British elasticity higher than 0.3.

This matter of demand elasticities is extremely important in determining how the world economy reacts to deficits in countries' payments balances and consequent excess supplies of their currencies. The world has operated under several sets of institutional arrangements for settling deficits and surpluses. Under all such arrangements, including the present system of flexible exchange rates, a deficit leads to a fall in the relative prices of goods produced in the deficit country, either because domestic price levels change or because exchange rates are free to respond to supply of and demand for currencies. If the price elasticities of demand for that country's imports and exports are sufficiently large (i.e., they sum to more than 1), the drop in relative prices will increase the demand for the country's currency relative to the supply and will correct the deficit. But if demands are inelastic, the fall in relative prices will reduce demand for the currency more than it reduces supply, making the deficit worse. You can see that this second reaction—a deficit producing a price reaction leading to a larger deficit—can be a major source of instability in the system of international financial arrangements.

Statistical studies of demand elasticities in international trade are not very conclusive. Most economists seem to think that demand elasticities are fairly large in the long run but much less so in the short run. American reactions to OPEC oil price increases seem to bear this out. The initial reaction to each of the major increases was for people and firms to dig deeper into their pockets and pay the price. Gradually, however, people learned to drive less and to keep their buildings cooler. Fuel-efficient cars, industrial equipment, and buildings replaced older, less efficient energy users. Alternative energy sources are continually being developed, and someday they may replace oil on a massive scale. Thus, although the initial demand elasticity was nearly zero, it may someday creep up past 1, so that the higher dollar prices of oil lead to a smaller, not a larger, dollar volume of imports.

We will see how world trade adjusts to surpluses and deficits in a few pages. First, however, we must look at capital-accounts payments, since they are another important part of the system of financial links among countries.

The capital account

The capital account is the flow of lending and investment from one country to another. Individuals and firms deposit their funds in foreign financial institutions, buy foreign securities, or invest directly in foreign subsidiaries for much the same reasons they make similar investments at home: They want to make a good return on their funds. Most of us never think of investing abroad because we have neither the knowledge nor the imagination to do so. But sophisticated individuals, financial intermediaries, and industrial firms move funds back and forth across international borders the .way some of us juggle our checking accounts. Whenever they think they can make a higher return abroad than they can at home, they sell assets at home and buy abroad. In some cases, this means withdrawing funds from Chase Manhattan and depositing them in the Deutschebank. In other cases, it means financing a new plant in Ontario and closing one in Detroit, or opening a Volkswagen plant in Bedford, Pennsylvania, rather than expanding the home factory in Wolfsburg, Germany.

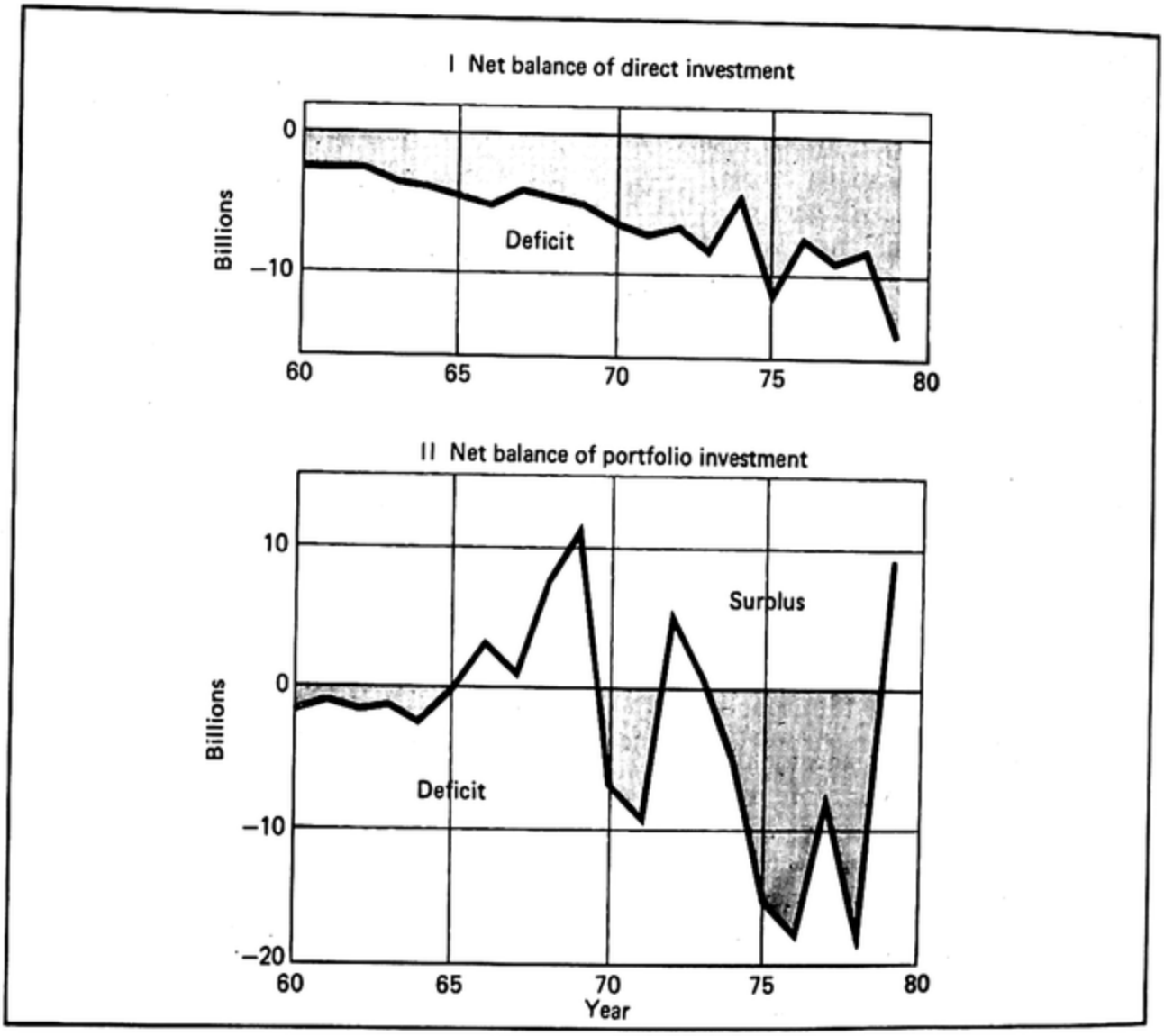


Figure 5 The U.S. capital-account balance 1960–1980 (in billions)

The net balance of direct investment was consistently negative from 1960 to 1980, as American firms invested more in their foreign subsidiaries than foreign firms invested in their American subsidiaries. The balance of portfolio investment (securities and deposits in financial intermediaries) was sometimes positive—when Americans and foreigners moved funds into this country from elsewhere—and sometimes negative—when Americans and foreigners moved their funds out of this country.

Source: *Survey of Current Business.*

To see why the capital account moves the way it does, it is helpful to break it down into two components. One is the balance of ***direct investment***—the investment that firms in one country make in subsidiaries located in another. The other component is the balance of ***portfolio investment***—the security purchases and deposits that individuals and institutions of one country make in another. Both direct and portfolio investment are transfers of funds, not physical investment in plant and equipment. Of course, these funds can be and often are used to finance physical capital formation.

The recent history of these two com-

ponents of the U.S. capital account is shown in the two panels of Figure 5. In each case, a net capital outflow—which represents an excess supply of dollars—appears as a negative figure, or deficit. The balance of direct investment shows up as a deficit throughout the 21-year period, since in every year between 1960 and 1980, American firms invested more in their foreign subsidiaries than foreign firms invested in their American subsidiaries. The balance of portfolio investment gyrated violently during the same period. In some years, both American and foreign investors moved their funds out of the United States and made deposits or bought securities elsewhere, where they expected a higher return. In other years, funds moved into the United States, as both American and foreign investors found higher yields here than elsewhere.

Most international investments affect the balance of payments for years to come, since most are profitable. Volkswagen built its plant in Pennsylvania to make profits, not to make cars. The plant's profit is the property of its German owners. It shows up in the U.S. current account as an outflow of funds from the United States to Germany, to pay for the services provided to the U.S. economy by the German-owned plant. Similarly, when an American investor buys shares in EMI (a British electronics firm), she expects dividends in return. If she gets them, they show up in the U.S. current account as an inflow of funds in return for the services provided to the British economy by her share of its assets.

The relationship between capital flows and the income that results from them is tricky, since the current-account income flow depends on the entire accumulation of past investment, and not simply on the current investment flow that appears in the capital account. Some historians used to believe in a characteristic pattern that

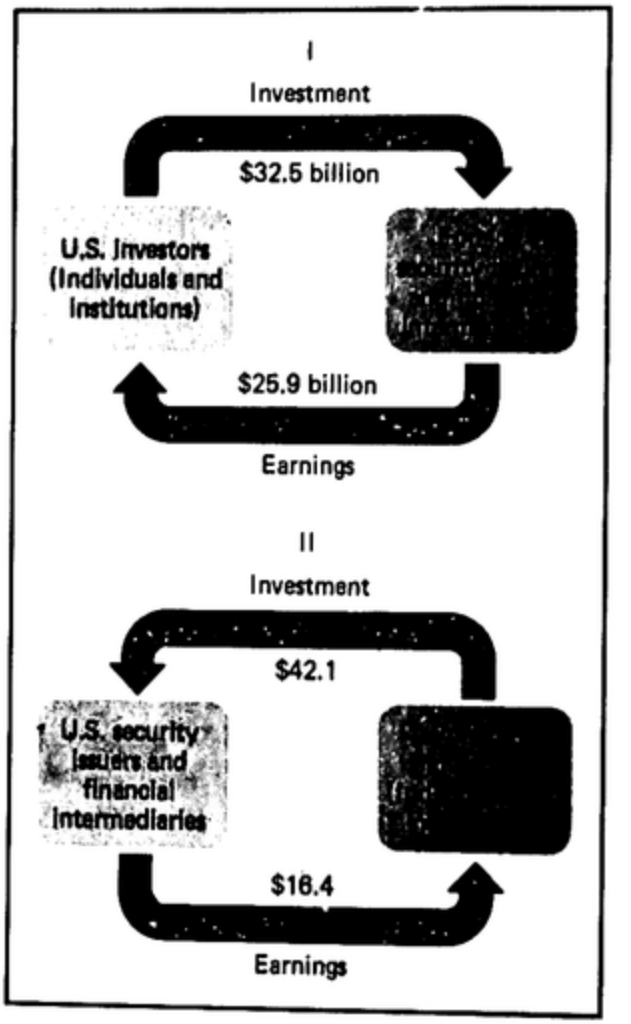


Figure 6 Portfolio investment and the balance of payments 1979 (in billions)

In 1979, Americans made $32.5 billion in portfolio investments abroad. Foreigners made $42.1 billion in portfolio investments in the United States, so that there was a net inflow of $9.6 billion from this portion of the capital account. In the same year, Americans received $25.9 billion in income on their past portfolio investments, and paid out $16.4 billion in earnings on past portfolio investments made by foreigners in U.S. securities and bank accounts. Thus, the investment income portion of the current account showed a $9.5 billion surplus.

Source: *Survey of Current Business.*

reflected the varying stages of development among countries. A relatively undeveloped country will have a capital inflow. At first, the inflow exceeds the income paid out to past investors. At this stage, the country is sometimes known as an *immature debtor.* As the accumulation of foreign

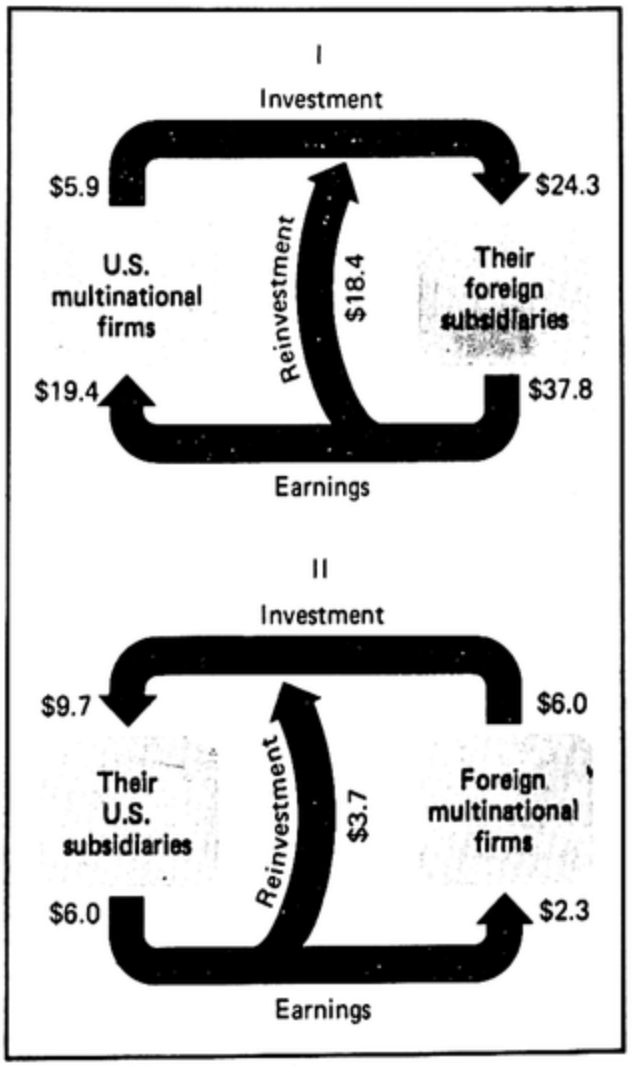


Figure 7 Direct investment and the balance of payments 1979 (in billions)

In 1979, U.S. multinational firms invested $24.3 billion abroad in their foreign subsidiaries. Most of this ($18.4 billion) was a reinvestment of earnings on past investment. Besides the $18.4 billion reinvested, however, $19.4 billion in earnings came back to the United States.

Foreign firms invested $9.7 billion in their U.S. subsidiaries in the same year. Only about a third of this ($3.7 billion) was a reinvestment of earnings on past investment; the rest came from the parent firms themselves. Unlike the American multinationals, foreign firms received repatriated earnings that fell far short of the $6.0 billion they directly invested in their American subsidiaries.

In each of the panels of the diagram, the total investment is a part of the capital account and the total earnings a part of the current account.

Source: *Survey of Current Business.*

investment builds up and the rate of inflow slows down, the income paid out as profit on past investment begins to exceed the new investment coming in, and the country becomes a *mature debtor.* At some point, the country may become developed enough relative to the rest of the world to begin to be a net exporter of capital. At first, the outflow of investment exceeds the inflow of income. At this point, it is an *immature creditor.* Eventually, however, the accumulation of past investment builds up relative to the current outflow. At this final stage, income returning from past investment begins to exceed its new foreign investment, and the country becomes a *mature creditor.*

U.S. foreign investment in a recent year relative to income on past investment is shown in Figures 6 and 7. Look first at Figure 6 (page 749), which pertains to portfolio investment. The first thing you will note is that the United States in 1979 was both a substantial exporter and a substantial importer of capital. On balance, foreigners invested about $9.6 billion more in the United States than Americans invested elsewhere. The flow of earnings to the United States exceeded the outflow of earnings, so that the current account showed a $9.5 billion surplus from the return on past investment. Thus, the United States was both a borrower (with a net inflow of investment) and a creditor (with a net inflow of income)! So much for the simple theory of stages.

Now look at Figure 7, which covers direct investment. Panel I shows that U.S. multinational firms invested $24.3 billion abroad in 1979, and received $37.8 billion in income from their foreign subsidiaries. Most of the investment was covered by reinvested foreign earnings, so that the $19.4 billion in earnings flowing into the country greatly exceeded the $5.9 billion that U.S. firms sent abroad. Thus, in its di-

rect investment, the United States looks like a mature creditor.

Or does it look like a mature creditor? Foreign multinationals were investing $9.7 billion in the United States during the same year and receiving only $6.0 billion in earnings on past investment. Thus, the foreign-owned sector of U.S. business was an immature debtor sector.

All these conflicting crosscurrents of investment merely underscore the fact that the foreign investment picture is far more complicated than it once was. Now that capitalist countries invest mainly in one another, direct and portfolio investments go both ways across national boundaries.

What motivates this great flow of finance capital, apart from the obvious fact that its owners are seeking high returns? The question requires two answers, one for portfolio investment and one for direct investment. Turn your attention first to direct investment.

Direct Investment

Investment in a foreign subsidiary is undertaken by firms that have been around, and expect to be around, for a long time. Such firms have *growth strategies*—long-run plans about their corporate futures. These strategies spell out what products the firm is likely to be producing over the next few decades, where their inputs are likely to come from, where these inputs can most profitably be combined to produce outputs, and where the markets for these products are likely to be over the time span of the plan. Products that require bulky inputs located in only a few countries are most profitably produced close to the sources of input supply. Products whose production is very labor intensive are most profitably produced in countries where labor is plentiful relative to demand, and unions are weak. Products whose markets are protected by tariffs are most profitably produced inside the tariff barriers, not outside. Products whose production is dangerous to the workers or to the regions around the factories are most profitably produced in countries that place a small value on worker safety or environmental protection. And countries that are so poor that they will fight one another with tax concessions to attract capital are profitable to almost any kind of business.

Most socialists who complain loudly about the American economy have at one time or another been confronted with the question: "If you don't like it here, why don't you move to Russia?" Few of them do, because they know they wouldn't like it there either. But ask the same question of the president of a large capitalist corporation, and he might well say, "That's not a bad idea. If they would only build us a big enough dock complex near Leningrad and guarantee a long-term lease, we could probably work out a deal." This is how multinational corporations operate. Their mode of conducting international business is changing the world economy at a faster pace than most of us can keep up with.

Multinational firms invest where it is most profitable for them, not where their owners happen to have grown up. For a long time, most direct investment financed projects in underdeveloped countries, but today the enterprises of the advanced nations are investing in one another's countries, pursuing a strategy of locating near their major markets. American business has been a leader in this kind of investment, but increasingly European and Japanese firms are building facilities in the United States itself. During the 1970s, direct investment in the United States increased tenfold, and the United States may again become a net importer of capital, as it was in the 19th century. This could hap-

pen if foreign firms decide it would be better to produce for the American market in America rather than producing for export to this country.

Portfolio investment

Since direct investment is geared to long-run growth strategies, it is not very sensitive to cyclical developments. Portfolio investment is quite sensitive, however. During the period of exceptionally high interest rates in the 1970s and early 1980s, the kind of "hot money" that is moved around in response to interest rate differentials became increasingly international in its travels. Much of it consisted of the accumulated profits of those who benefited from OPEC. Its owners held a share of their newfound wealth in very liquid form and were quick to transfer it from one country to another in response to cyclical changes in interest rates. If you look at Figure 5, you can see that the amplitude of the swings in portfolio investment grew enormously over this period.

Oil millionaires are happy to move funds to the United States in response to high interest rates. So are many other institutions and investors, both American and foreign. They will move their funds away from the United States when it looks as though they can make a greater gain elsewhere. Their willingness to move funds does a lot to equalize interest rates throughout the world's financial markets.

But whenever there is a business cycle change in one of the major countries, a lot of money is set in motion. If Italy enters into a recession while the rest of the capitalist world is prosperous, Italian interest rates will drop. This will lead to a capital flight from Italy, and an excess supply of Italian currency, the lira. If there is a sudden upsurge in inflationary expectations in the United States, U.S. interest rates will rise and capital will flow in.

If you ever try to make sense out of the figures on portfolio investment, you will have to remember that capital movements reflect mainly *changes* in interest rate differentials, not levels. If the U.S. interest rates rise relative to those in the rest of the world, the United States will have a sudden capital inflow and there will be an excess demand for the dollar. But once all the "hot money" has moved, it will stay put until there is another disturbance of interest rates. Thus, a country will characteristically suffer a portfolio capital outflow at the start of a recession, when its interest rate has just dropped relative to the average rates elsewhere. Then, the capital account will stabilize until a recovery gets under way. At this point, a rise in the interest rate relative to those abroad will signal the beginning of a capital inflow. Again, once the funds have moved, the capital inflow will be over until the next disturbance in the world pattern of interest rates. This disturbance need not be of domestic origin. If the major countries of Europe simultaneously go into a slump, capital will move to the United States. When they recover, it will flow back.

During the general elections in Canada in the summer of 1980, the National Democratic Party candidate, Mr. Broadbent, campaigned on a platform calling for "an interest rate policy made in Canada." This must have embarrassed those of his supporters who understood international finance. The money markets of the United States and Canada are so closely intertwined that any attempt by the Canadians to keep their interest rates below those in the United States would result in a massive capital flight and excess supply of the Canadian dollar. The resulting strain on the balance of payments would quickly force the Canadians to give up their attempt to go it alone.

Now that you have some understanding of how both the capital and the current

accounts of individual countries fluctuate, you are ready to see how the whole system of trade and finance fits together, and how the demands for and supplies of the various world currencies are brought into line with one another. That is the subject of the next major section.

The world payments system

The demand for and supply of each of the world's currencies depends, as you know, on the prices, income levels, interest rates, and prospective yields on long-term investments throughout the world. Sometimes the pattern of these forces produces an excess demand for a particular currency and sometimes an excess supply. The world payments system, if it is to function smoothly, must not only enable one country's currency to be converted into another, but must also provide some mechanism for handling temporary demand-supply imbalances and eventually eliminating them. This section surveys how temporary imbalances and adjustments to equilibrium have been handled in three historical periods: (1) that of the ***gold standard,*** which functioned in the 19th and early 20th centuries, until it collapsed under the strains of two world wars and a worldwide depression; (2) that of the ***Bretton Woods system,*** which functioned from the end of World War II until its breakdown in the early 1970s; and (3) that of the ***floating exchange rates system,*** which has been in operation since 1973.

The gold standard

To understand how world payments systems work, it helps to begin with the tale of a mythical world, in which countries use gold coins for currency. In each country, the markings on the coins reflect that country's history, politics, and language. But since coins can be melted down and recast quite cheaply, currencies can, if necessary, be converted from one nation's to another's in the goldsmith's cauldron. People who mine gold can turn it into money at the same goldsmiths' establishments, and those who want gold jewelry (or reflective garments for space travel) can melt their money down to create beautiful (or useful) objects.

In such an age of gold, what would happen when a citizen of one country wanted to buy wheat or common stock from the citizens of another? Obviously, the transaction would be made in gold, since gold would make up a common world currency, acceptable everywhere. Thus, gold would be the universal *medium of international transactions.*

Suppose that in such a world Americans wanted to buy British goods and assets worth more in terms of gold than those the British wanted to buy here. This might happen, for example, because the gold prices of British goods were attractively low relative to gold prices of American substitutes. The amount of American currency seeking British goods would be greater than the amount of British currency seeking American goods, so that there would be an excess supply of American currency. But unlike paper dollars, gold dollars would be acceptable to the British, who could melt them down and convert them into any gold currency in the world. In effect, the entire U.S. money supply would function as *international reserves,* an acceptable international payments medium to tide the United States over its payments imbalance.

This could not go on forever, of course, since a hemorrhage of gold would deplete the U.S. domestic money supply and raise the money supply in Britain. The initial reaction to these money supply changes would be a rise in U.S. interest rates (because of "tighter" money) and a fall in

British interest rates. There would be a temporary capital flow to the United States, more or less offsetting the current-account deficit. But as long as American prices remained out of line with British prices, there would be a current-account deficit, though the capital-account inflow would only be temporary.

The combination of a dwindling money supply, high interest rates, and negative net exports would eventually produce a recession in the United States. One of the effects of the recession would be a drop in U.S. imports, caused by the decline in GNP. Like the higher interest rates, the cyclical drop in imports would offset the effects of the price imbalance. But this effect would only be transitory. As long as American prices were too high, the United States would lose gold during prosperous times and the U.K. would gain gold. The inevitable long-term result of the differences in monetary growth between the two countries would be more inflation in Britain than in the United States: Britain's prices would rise relative to America's prices. Assuming that the demand elasticities were big enough, this change in relative prices would correct the fundamental cause of the gold movement and restore equilibrium.

You can see that a world with a universal currency (gold) has several ways of adjusting to payments imbalances between countries:

1. The universal gold currency provides every country with a source of *reserves* to meet temporary imbalances in its international payments.
2. Because the reserves are, in fact, the domestic money supply, a loss in reserves triggers off immediate *cyclical changes* in capital flows and imports that limit the reserve loss.
3. Since the price level follows the money supply in the long run, a gold outflow leads to *price changes* that eventually end it.

At no time in recent centuries has the world trade and payments system been conducted with a common world currency. But throughout much of the 19th and early 20th centuries, trade among the world's major nations was regulated by a payments system that in most respects was like the mythical age of gold. Although countries all had money supplies made up of paper curreny, bank deposits, and silver as well as gold, their state treasuries and banks were committed to converting other forms of money into gold at a fixed price. This meant that a paper dollar was effectively title to a certain quantity of gold. It was also title to a certain quantity of pounds, purchasable with that gold. If exchange rates got out of line with gold values of the corresponding currencies, arbitrage would take place until they were back in line. This fixed the relative values of different currencies in terms of one another. So effective was this link that most transactions were conducted in a single currency, the British pound, that could readily be converted to gold or to any other gold-linked currency.

Another similarity between the mythical age of gold and the historical era of the gold standard was the close connection that most countries maintained between their total supplies of domestic money and their holdings of gold. Although any country's money supply was larger than the value of its monetary gold, there existed a rough proportionality between the two. In many places, banks kept gold as a reserve against their deposit liabilities. In the early days of the Federal Reserve System, the amount of money created by the Fed and its member banks was tied (in a complicated way) to the Treasury's holdings of monetary gold. This proportionality meant that whenever a country had a gold out-

flow or inflow connected with a price disequilibrium, its domestic money supply would change to restore equilibrium.

The gold standard had its faults, however. They often go unrecognized by people who long to return to a world monetary system in which every currency is "backed by gold." The main fault was the inflexibility of the supply of world money. Prospecting for gold was influenced by the demand for gold, but prospecting does not always pay off. For long periods of time, the world money supply would change very little. This had a deflationary impact on prices and production, as the demand for money tended systematically to outstrip the supply. Suddenly there would be a major gold strike, like those in California, Alaska, and South Africa and there would be a worldwide inflation.

Because of uncertainties in the growth of the world gold supply and the resulting failure of gold to keep pace with the growth in world trade, countries felt obliged to expand their money supplies relative to their gold holdings, so that gold became an ever-dwindling fraction of the world money supply. Once gold had become a relatively small part of a country's money supply, most of that supply was useless as a *reserve* against balance-of-payments deficits. When such a country got into payments difficulties, its banks and treasuries would start to run out of gold. In the end, their governments would have to declare a "cessation of convertibility." The domestic currency could no longer be converted into gold because the cupboard was bare. This happened to the dollar during the U.S. Civil War (1861–1865) and to most of Europe's currencies during World War I (1914–1918). Restoration of convertibility after World War I was difficult because of the political weakness of many governments. Hardly had convertibility been restored throughout much of the world before the payments system was buffeted by the Great Depression (1929–1941) and World War II (1939–1945), which spelled the end of the gold standard. Like the Roman Empire, the gold standard died slowly and painfully. It was the victim of its fundamental weakness—an inability to provide sufficient reserves for an unstable world.

The Bretton Woods system

The collapse of the gold standard system between 1914 and 1945 had serious consequences for prosperity and growth throughout much of the world. When countries could no longer redeem their currencies with gold, they were forced to curtail imports abruptly by quotas, tariffs, and other forms of direct control, and by devaluing their currencies relative to gold in a vain attempt to attract exports from one another. When all countries pursued such policies, the results were a shrinking of trade and a loss of the benefits of international specialization and the pursuit of comparative advantage. Such policies—often referred to as "beggar my neighbor" policies—were mutually destructive. But individual countries found it difficult to resist retaliating when they were confronted by tariff walls and currency devaluations by others.

Toward the end of World War II, the major countries negotiated a restructuring of the world monetary system. The outcome was the Articles of Agreement of the International Monetary Fund, more commonly called the *Bretton Woods Agreement*, after the place where it was signed. This drew up the rules under which world monetary affairs were conducted until a new arrangement was worked out between 1971 and 1973. The main articles of the Bretton Woods Agreement were the following:

1. *Exchange rates* among currencies were to be fixed. Each participating country

was required to establish a *par value* for its currency in terms of gold. Under normal circumstances, the gold parity would be maintained, and so would the exchange rate with other currencies tied to gold. Only a "fundamental disequilibrium" in a country's balance of payments would justify a change in the par value of its currency.

2. A pool of various currencies was set up —the ***International Monetary Fund,*** or ***IMF.*** Countries subscribing to the pool were required to deposit a *quota* of gold, dollars, and their own currencies. Members that were faced with temporary payments deficits could borrow limited amounts of any currency from the fund to tide them over, but had to repay in currencies freely *convertible* with gold.

The fixity of par values for currencies in terms of gold made the Bretton Woods system look like the gold standard in one important respect. Under ordinary circumstances, the parties to a contract could count on exchange rates not to change over the life of the contract. It was widely believed that this arrangement encouraged trade and international investment, since it provided one element of certainty in an arena with many uncertainties.

Two important elements of the gold standard were missing, however. One, governments were not obliged to buy and sell gold in exchange for their currencies. Some immediately announced *covertibility with gold,* many did after a while, and some of the framers of the agreement hoped all would eventually. But in the immediate postwar period, many governments controlled their exchange transactions very carefully. Two, domestic money supplies were not closely tied to their countries' gold holdings. Thus, the adjustment processes built into the gold standard system—the means for correcting payments imbalances—were absent from the Bretton Woods Agreement.

How did the system established by the Bretton Woods conference function? How was one country's currency *converted* into that of another? How were *reserves* supplied to tide countries over temporary imbalances in their payments? And what was the *equilibrium* mechanism that brought prices into line to eliminate persistent imbalances in the demand for and supply of various currencies?

First, the matter of conversion of currencies: This was largely accomplished on currency exchanges of the sort discussed early in the chapter. Because exchange rates between currencies were fixed by their par values, the exchanges did not set prices, but merely made transactions. They were willing to do this because they collected commissions from every exchange and because the central banks of the countries whose currencies were traded on the exchange were pledged to purchase any excess supplies with other currencies or gold. These same central banks would meet excess demands for their currencies by exchanging them for other currencies or gold.

The central banks and their affiliated national treasuries were responsible for balancing supplies of and demands for their currencies. To do this, the central banks had to hold inventories to cover day-to-day fluctuations in demand and supply. A central bank has no trouble handling an excess demand for its own currency, since it is the producer of that currency. To deal with excess supply, however, it must have *reserves* of other assets that foreigners will accept. Under the Bretton Woods system, as under the gold standard, gold functioned as one such asset. So did official holdings of currencies convertible into

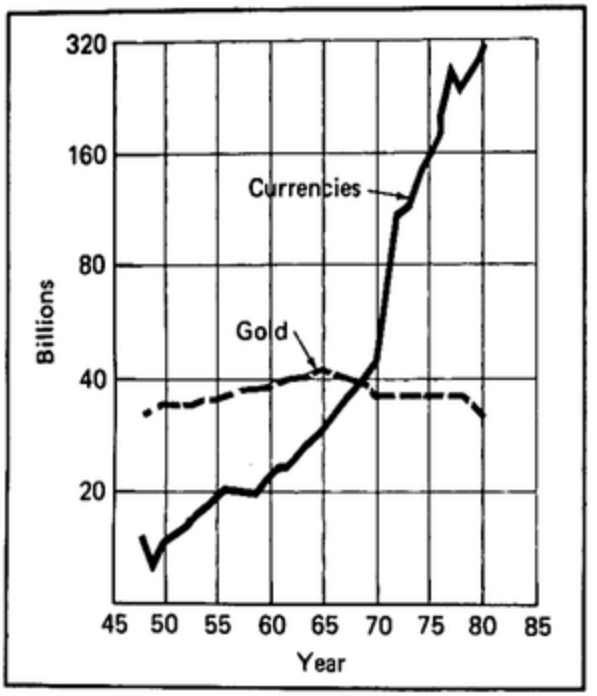


Figure 8 World official reserves of gold and currencies 1948–1980 (in billions of SDR units)

Since World War II, world gold reserves have not increased, but currency reserves have grown enormously.

Source: International Monetary Fund.

gold. As the system matured, official holdings of these *convertible currencies* grew much faster than the world supply of monetary gold (see Figure 8). The U.S. dollar soon became such a dominant *reserve currency* that the Bretton Woods system could justifiably have been called the dollar standard.

You will remember that the early part of this chapter explained the structure of the U.S. balance-of-payments accounts. Later sections developed more fully the economic factors lying behind the balances in the current and capital accounts, but little was said about the *official settlements* account. These official settlements are the transactions by which governments intervene in the markets to absorb excess supplies of and satisfy excess demands for their own currencies. As they do so, they lose or acquire international reserves.

The framers of the Bretton Woods Agreement were well aware of the problems that inadequate reserves created for the gold standard system. This was why the International Monetary Fund was established in 1945. Its lending activities were intended to "bail out" countries whose reserves had reached critically low levels. *Quotas* of gold, U.S. dollars, and national currencies were deposited by the members in rough proportion to the size of their economies. (The U.S. quota, which was by far the largest, was 23 percent of the total subscription.) Members could borrow from the IMF in limited amounts to supplement their own reserves. In 1967, the members of the IMF agreed to create a new international asset, the Special Drawing Right (or *SDR*). The SDRs were distributed to countries in proportion to their quotas. The member governments agreed to accept limited quantities of SDRs in exchange for their own currencies. Thus, the SDR constituted a new international reserve, along with gold, convertible currencies, and ordinary borrowing rights at the IMF.

Essentially, the IMF is a mutual aid society, created by members because they had learned from experience that their individual fortunes depended on their collective solvency. Whatever dreams its founders may have had about its becoming the central bank of the commercial world, however, have been dispelled by the unwillingness of member countries to give up too much of their sovereignty. Reserves created through the IMF have never reached as much as 20 percent of the world's total. Owing to its members' timidity in expanding the IMF's role in the world, the U.S. dollar grew to dominate world reserves. Although this worked out

well for a time, it eventually meant the doom of the Bretton Woods system.

Payments equilibrium under the Bretton Woods system

If all countries had tied their domestic money supplies to their international reserve positions, the Bretton Woods system would have functioned much like the gold standard. A reserve loss would have caused a monetary contraction, high interest rates, capital inflow, a recession with a drop in imports, a fall in prices relative to those elsewhere, and an eventual restoration of balance. A reserve gain would have triggered a monetary expansion, a drop in interest rates, a capital outflow, a boom with rising imports, and a rise in prices relative to those elsewhere. Again, the restoration of equilibrium would have been left to more or less automatic forces.

But the Bretton Woods conference was greatly influenced by the thinking (and presence) of John Maynard Keynes. Much of his life's work had been devoted to understanding what went wrong with the world capitalist system between 1914 and 1945. One of his conclusions was that the gold standard was fatally flawed by its inability to provide adequate reserves. Another was that countries had the power to control domestic prosperity by managing their budgets and money supplies according to what we now call Keynesian principles. Submitting to the automatic "discipline" of the gold standard was not consistent with managing the domestic economy.

Countries that want to manage their domestic economies must somehow divorce their domestic monetary changes from their international reserve positions. This means, for example, that they must not allow changes in official holdings of gold and foreign exchange to alter their domestic banks reserves. During the 1960s, Germany had considerable difficulty in managing its money supply, since its payments surplus fueled rapid growth in the reserves of its banking system. Germans were earning foreign currencies and converting them to marks at their banks. The banks would present these currencies to the Bundesbank (central bank) in exchange for reserves, and the German monetary base would expand as a consequence. This made it difficult for the German government to maintain price stability in the face of its payment surplus.

Countries that successfully controlled their money supplies still somehow had to contain changes in their international reserves. A surplus was relatively painless, since an expansion of international reserves caused no particular problems for the country that was gaining them. But a deficit country with limited reserves had to take action. It had to use its domestic monetary and fiscal measures to stop its reserve loss. This meant tightening money, raising interest rates to encourage a capital inflow, cutting the budget, and slowing the growth of GNP to reduce imports. Eventually, it had to control the growth of prices. To a degree, it had to act as though it were subject to the automatic discipline of the gold standard. The discipline imposed by limited reserves was one-sided, however. Countries that were running a surplus did not *have* to expand demand and raise their rates of price increase. This led some analysts to complain that the Bretton Woods system had a *deflationary bias:* It forced some countries to deflate without forcing others to inflate. Although this may have helped control world inflation, it also raised the world unemployment rate.

In some cases, it didn't even work to bring the payments of deficit countries into balance. Currencies that were in seri-

ous excess supply did not respond quickly to doses of deflationary medicine. Remember that demand elasticities in international trade are fairly small in the short run. Changes in the domestic price level relative to those elsewhere may have little impact on the excess supply of the domestic currency for several years. (In the extreme case, it may even increase it for a while.) The Bretton Woods Agreement foresaw the possibility that a currency might become so *overvalued* that the only practical way to restore payments balance was through *devaluation*, a unilateral decision by a government to lower the par value of its currency in terms of gold and, therefore, in terms of every other currency. Spared the agony of a long domestic deflation, with high unemployment, a country could with one stroke lower the relative price of its goods on domestic and world markets. Although the favorable effects of the price change required time to be felt, the change itself was quick.

Devaluations were not painless, however. Governments were reluctant to devalue because devaluation was an open admission that other policies had failed, and because it also immediately raised the prices of imports. In Britain, which devalued twice (in 1949 and 1967), the immediate impact was a cut in real wages, since so much of Britain's food has to be imported.

Under the Bretton Woods system, therefore, devaluations were rare, particularly for major currencies. But occasionally they became inevitable. The mere fact that a currency was in trouble because its country was rapidly losing reserves often became the center of attention of the international currency market. You may know the saying, "Where the carcass is, there will the eagles gather." With a currency, the carrion eaters were not eagles, but speculators, both domestic and foreign, who unloaded a currency they thought might be devalued. This increased the excess supply of that currency and could force a devaluaton by worsening the reserve position of the country whose currency was being speculated against.

The demise of the Bretton Woods system

The system finally collapsed when the dollar itself came under heavy speculative attacks in the early 1970s. For more than two decades, the world had acquired enormous quantities of liquid assets in the United States—nearly $100 billion by 1970. During the final years of the Vietnam War, high interest rates in this country attracted enough short-term foreign capital to offset the deterioration in the current account brought on by inflation. But when the winding down of this conflict brought on the recession of 1970–1971, U.S. interest rates dropped, and there was a massive capital outflow. The Treasury lost about a third of its gold holdings, and the holders of dollars became jittery about the prospect of devaluation.

You probably have had the experience of eating too much rich food, perhaps on Thanksgiving. It all tastes so good as you pile it in. Then, suddenly, one bite is too much, and you begin to regret deeply all that you have eaten over the last half hour or so. Something like this happened to dollar holders in the early 1970s. The dollar had been relatively scarce in the late 1940s and 1950s. In the jargon of international finance, it was *undervalued* relative to other currencies, making American goods a bargain on world markets. The United States consistently ran a trade surplus, and only a large outflow of foreign aid and private long-term capital allowed the rest of the world to acquire the dollar-denominated assets that it wanted. But the changes that took place during the Vietnam War weak-

ened the purchasing power of the dollar relative to that of other currencies, so that it came to be *overvalued.* An undervalued currency is a very attractive asset to hold because it is sure not to be devalued. But an overvalued currency is a different matter entirely.

The position of the dollar was made more precarious by the existence of *Eurodollars,* which are bank deposits in other countries denominated in terms of U.S. dollars. The world financial community had created its own annex to the U.S. banking system, entirely outside the control of the Federal Reserve or any other U.S. governmental institution. Although Eurodollar deposits are not legally part of the U.S. money supply, their function in the structure of the world asset holdings is much like that of U.S. bank deposits. If the holders of Eurodollars decide they would rather hold marks, yen, or pounds, their attempt to convert their holdings into these other currencies creates an excess supply of dollars. The mass of dollar-denominated assets whose owners might decide to look for a sounder currency is swollen by the Eurodollar.

Through a succession of actions taken over the 1971–1973 period, the major financial powers finally abandoned their efforts to maintain the par value of the dollar. It was allowed to fluctuate in value according to the dictates of supply and demand. Once the world's major reserve asset was cut loose from its Bretton Woods par value, so were all other currencies, and the world payments structure shifted over to a system of *fluctuating exchange rates.*

The world system of exchange rates that has regulated international finance since 1973 is really a hybrid, with elements of the fixed-rate system grafted onto an exchange market in which relative prices of various currencies are free to change in response to supply and demand.

In a *pure system* of floating exchange rates, governments would stay out of the exchange markets altogether, except when they needed foreign exchange to make loans or purchases abroad. They would never try to stabilize the values of their currencies. If the United States were running a deficit in its combined current and capital accounts, neither the Federal Reserve nor any foreign central bank would intervene in the dollar market. There would be no official settlements transactions in the balance of payments. The excess supply of the dollar would be permitted to depress its value relative to that of currencies in excess demand. Dollar goods would become relatively cheaper, and goods priced in other currencies relatively dearer. With appropriately large demand elasticities, the fall in the value of the dollar would raise the demand for American goods and, therefore, American dollars sufficiently to eliminate the deficit.

All of this would take place by way of the currency market itself; no internal price adjustments would be necessary. Recall that under both the gold standard and the Bretton Woods system, the normal processes of adjustment to payments disequilibriums required changes in domestic price levels. The values of currencies in terms of one another were fixed. Their values in terms of goods could therefore be kept in line only by a system that equalized inflation rates in different countries. The pure system of floating exchange rates frees domestic stabilization policy from the sometimes harsh discipline of the balance of payments. If a country with a payments surplus wants price stability, it need not inflate to eliminate the surplus. Its currency will appreciate on the world market, making foreign goods cheaper in its domestic market, and the prices of its home-produced goods can remain stable. By the same token, a country with a persis-

tent deficit need not engineer a prolonged domestic recession just to bring its prices down relative to those elsewhere. This can be accomplished directly by a drop in the value of its currency on the exchanges of the world. And because governments never intervene in the currency markets, they don't have to keep reserves to tide them over short-term imbalances in the demands for and supplies of their currencies. Without the responsibility for their currencies, they have no need to hold gold or foreign currency balances. Under a pure system of floating exchange rates, neither ordinary nor special drawing rights at the IMF would serve any purpose either.

You can see why so many economists favor fluctuating exchange rates. Free-market advocates like Milton Friedman argued the merits of fluctuating exchange rates long before the overvaluation of the dollar forced the world to abandon the Bretton Woods system. Why, then, have the world's currencies historically been linked by fixed rather than fluctuating exchange rates, except in times of crisis? Why do many business people, government officials, and economists argue for a return to some system of fixed rates? And why do governments intervene in the markets to support their currencies even under the current system of floating exchange rates?

There are several reasons. First, many of the direct participants in the exchange markets seem to prefer fixed rates. It is less complicated to think ahead if the number of variables is smaller rather than larger. Second, the day-to-day role of speculators under a fixed exchange rate system is far smaller than it is under a flexible system. Except when a currency is in danger of devaluing (or revaluing, which means rising in par value), the speculators play a minor role in the fixed-rate system. But most important is the question of stability. Remember that in the short run demand elasticities are smaller than they are in the longer run. This means that a fall in the relative value of a currency may add to the excess supply rather than eliminate it. No one knows with confidence what the relevant elasticities really are, but it is at least arguable that a pure floating exchange rate system would be very unstable.

For these reasons, many governments try to stabilize the values of their currencies under the current exchange rate system. The system is sometimes called a *managed* or *dirty float*, in contrast to a pure float. Nonetheless, rates have been permitted to fluctuate rather widely. Figure 9 shows how the dollar has varied relative to other major currencies since 1973. The most interesting thing about the pattern of changes is its diversity. The dollar has risen a lot relative to the lira and fallen a lot relative to the mark. Without the freedom to fluctuate on the currency exchanges, the relative values of the dollar, lira, and mark would have been inconsistent with payments equilibrium during the 1970s, and either the domestic economies of the United States, Italy, and the German Federal Republic would have had to adjust to the problem somehow or there would have been devaluations and revaluations. The 1970s, with their enormous changes in the prices of food and energy, were bound to be turbulent. The flexibility of the managed float served the world well. Trade expanded enormously.

The times were less troubled than they might have been. In some respects, the world was lucky. If it had still been on the gold standard in 1973, it is hard to see how it could have handled the enormous changes in relative prices without a reversion to the "beggar-my-neighbor" tariffs and devaluations of the 1930s. The free market is a very imperfect institution, but fortunately it does respond quickly to major disruption.

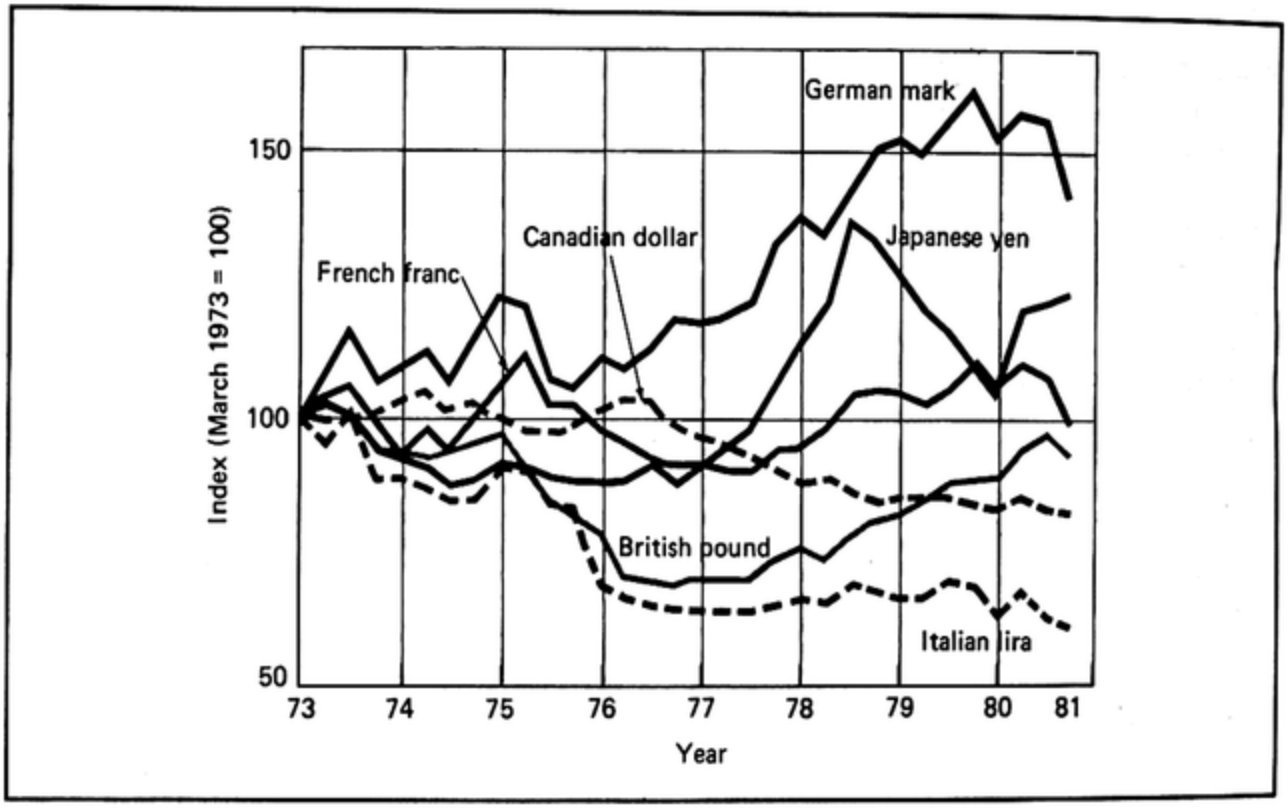


Figure 9 Dollar value of selected currencies 1973–1980

In the seven-year period following the floating of exchange rates, the mark steadily rose relative to the dollar, the lira steadily fell, and other currencies fluctuated.

Source: *Economic Report of the President.*

Summary

This chapter has focused on the monetary links among the world's trading nations. Here are some of its most important points:

1. Monetary exchanges between countries with different currencies require that buyers convert their currencies into those of sellers before the exchange transactions can be completed.
2. This conversion is accomplished on currency exchanges, which are markets on which currencies are bought and sold. The participants in these transactions are the buyers of goods and assets, financial institutions, speculators, and central banks.
3. Every country has a balance of payments, which is the balance of demands for and supplies of its currency on the currency exchanges. The balance of payments is made up of a current account (goods and services), a capital account (financial assets), and an official settlements account (government transactions in currencies and gold).
4. Changes in a country's current account are heavily influenced by its GNP and general price level relative to those in the countries with which it trades. Changes in its capital account reflect changes in its interest rates relative to those in other countries. Changes in its official settlements account reflect its government's efforts to stabilize the currency markets.

5. The world payments system is the set of international arrangements for settling payments among countries. The payments system provides a market for currency conversion, a supply of reserves to tide countries over temporary imbalances in their payments, and a set of economic mechanisms for correcting chronic payments imbalances. The oldest such system is the gold standard system. It collapsed during the period of the two World Wars and the Great Depression. It was succeeded by the Bretton Woods system, which foundered in the early 1970s and was replaced by the current system of floating or flexible exchange rates.

6. Under the gold standard system, world currencies were tied in value to gold, which made up a substantial portion of the world's money supply. A country that had a balance-of-payments deficit settled it by losing gold. This gold loss temporarily carried the country over its deficit, but the resulting money supply contraction caused unemployment and domestic deflation. The result was a decline in imports and an increase in exports, which corrected the deficit. The gold standard worked well enough in normal times, but it could not cope with the World Wars and the Depression, principally because the world's gold supply was insufficient to provide its countries with enough reserves to cover major payments imbalances.

7. The Bretton Woods system, which was instituted in 1945 by international agreement, was in some respects like the gold standard. Currency values were tied to gold, which was an acceptable means of international payment. But under the Bretton Woods system, U.S. dollar balances were the major form of world reserves. Governments were responsible for supporting the values of their currencies by buying excess supplies with their reserves of gold and dollars. They were also responsible for coping with persistent payments imbalances, either by means of a combination of fiscal and monetary policies, or, in extreme cases, by changing the gold values of their currencies. The Bretton Woods system worked tolerably well until the early 1970s, when doubts about the value of the dollar caused a world monetary crisis.

8. Since the early 1970s, the world payments system has been governed by a system of floating or flexible exchange rates. The relative values of currencies are free to fluctuate in response to supply and demand. In principle, governments need no longer hold reserves or attempt to stabilize the values of their currencies. If a currency is in excess supply, its value will drop and its country's goods will immediately become cheaper on world markets, stimulating the demand for the currency. In practice, most governments have intervened in the currency markets to offset speculative changes. But the flexible-rate system has so far proved able to adjust to enormous increases in the relative prices of energy and food without a breakdown.

Key concepts

Balance of payments
Currency exchanges
Exchange rates
Arbitrage
Speculation
Current account, capital account, and official settlements

Monetary reserve assets
Direct and portfolio investment
Gold standard, Bretton Woods (fixed exchange rates) system, floating exchange rates system
International Monetary Fund (IMF)

Questions for review

1. a. The United States and Britain are trading partners. What would happen in the currency market in dollars and pounds if real GNP in the U.S. increases, while in Britain it remains constant?
   b. If the rate of growth of real GNP is the same for both countries, can you assume an equal rate of growth of both countries' imports?
2. What is the difference between direct investment and portfolio investment? Are both equally sensitive to cyclical variations such as changes in interest rates?
3. How did the Bretton Woods system of international payments contribute to worldwide unemployment?
4. a. What are some arguments in favor of fluctuating exchange rates?
   b. If there are advantages to fluctuating exchange rates, why do governments try to stabilize the value of their currencies?

36

Economic Development

As you read and study this chapter, you will learn:

- what the economic characteristics of underdeveloped countries are
- why they are underdeveloped
- what policies may speed their development, and what some of the limitations on development are

Even if Christmas was not the big day of the year when you were a child, you probably remember "The Night Before Christmas":

> 'Twas the night before Christmas, when all through the house
> Not a creature was stirring—not even a mouse.

Then it went on about chimneys and children, sugarplums and Saint Nicholas. Maybe you even have it memorized.

Does it make you feel nostalgic? Probably not at your age. But be glad you grew up in a country where such a syrupy vision of childhood is possible.

Suppose you lived in Bangladesh, an Asian country of 90 million people, whose *per capita GNP was under $100 in 1979* (compared to $10,600 in the United States). Can you imagine what life must be like on $100 a year? No house, only a hut. No mice, but plenty of rats. No stockings, no chimney, no hopes of St. Nicholas, no beds. And no visions of sugarplums, either. Just hunger,

day after day after day. About 2,000 people in Bangladesh die from starvation and malnutrition every day, Christmas included.

Countries like Bangladesh are such an embarrassment that world officials never seem to know what to call them. At one time they were called "backward countries." This seemed condescending, particularly since so many of these areas have possessed very high cultures for thousands of years. "Underdeveloped" was not much better. Then the official term became "developing," until it was pointed out that in many such places no development was taking place. This was replaced by "less developed," which was studiedly noncommittal.

The governments in question think of their countries as composing the ***Third World*** (as distinct from the industrial-capitalist ***First World*** and the Soviet-bloc ***Second World).*** In this chapter, such countries will usually be called *Third World*, sometimes *less developed*, sometimes *developing*, and sometimes *underdeveloped.*

Of the 4 to 5 billion people living on the globe in 1980, about three fourths inhabited the Third World, including the People's Republic of China. Many of these people were not at all poor. But about half the world's population lived in countries whose GNP was under $400 per capita. Virtually all of these people lived in Asia or Africa.

Much of the interest in the Third World is humanitarian. You may well have been attracted to economics because you thought you could learn how to help people. Well, much of the Third World really needs help, if anyone does. But even people whose humanitarianism doesn't extend much beyond their immediate family and friends take an interest in the Third World. There are several reasons:

1. About one third of world merchandise trade involves Third World countries. They are major sources of primary products and consumers of manufactured goods. If their economies were better developed, they might provide cheaper sources of supply and better markets.

2. The industrialized countries have extended substantial amounts of development assistance to the Third World. For example, the (First World) members of the Organization for Economic Cooperation and Development gave about $25 billion in aid to Third World countries in 1980 alone. Private investors in the First World have also made large portfolio and direct investments in the Third World. These private capital movements from the First World to the Third outweigh official development assistance. Both public and private investors believe they have a substantial stake in the future of the Third World.

3. There is a lot of poverty close to home, in the Caribbean, Central America, and South America. Haiti is one of the poorest countries in the world, and most of Central America is also very poor. The political stability of the western hemisphere is closely linked to the pace of development in these areas.

4. Most important, by the year 2000, 80 percent of the world's population will live in the Third World. When world population stabilizes, as it must someday, the percentage will be still higher. The long-term future of humanity, if there is to be one, will be determined by what happens in the Third World.

This chapter is divided into three main parts. The first surveys the characteristics of Third World countries and compares them with First World countries. The second explores why the Third World is so poor. The third outlines the various development strategies that might successfully get Third World countries out of their current state of relative poverty.

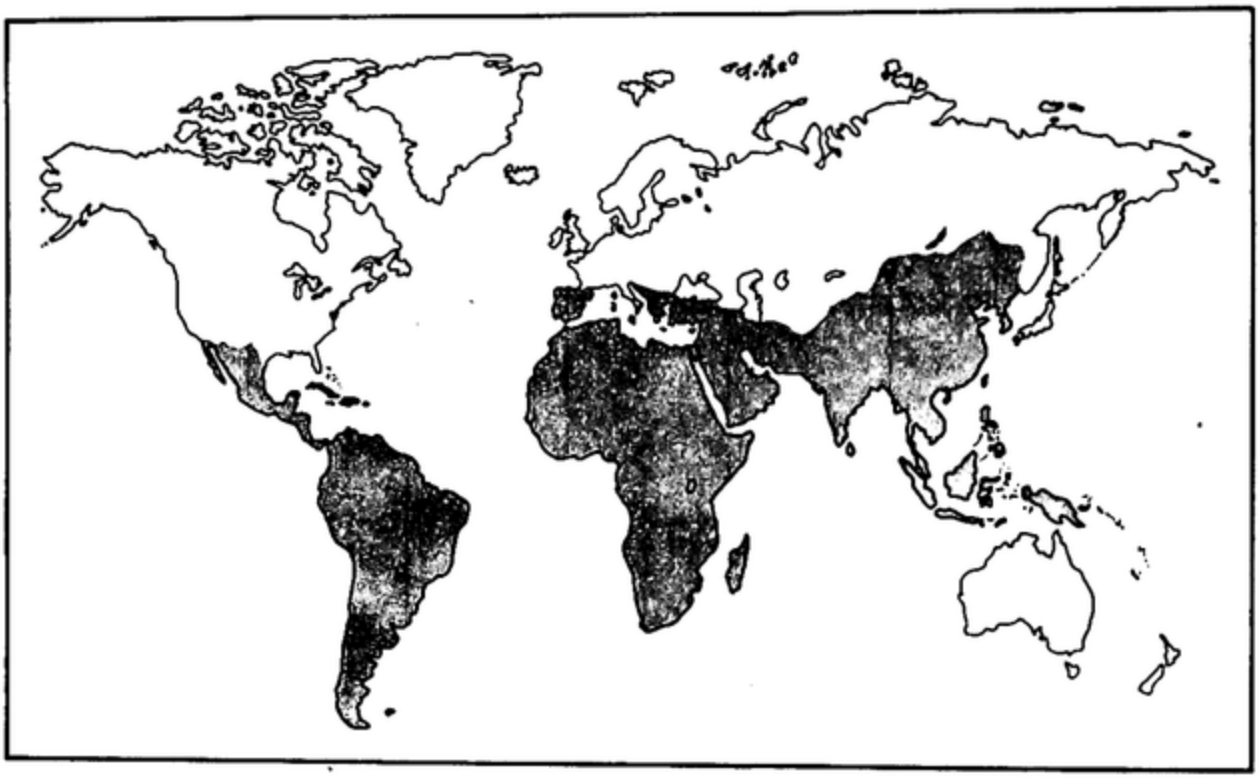


Figure 1 The Third World

What are the Third World economies like?

Tolstoy begins his great novel *Anna Karenina* with the words "Happy families are all alike; every unhappy family is unhappy in its own way,"—an exaggeration, of course, but one that is full of insight. If you were to visit first the 18 countries that the World Bank classifies as "industrial market economies," and then the 96 countries it classifies as low or middle income, you would certainly find the industrial countries more like one another than the others are. It is a lot easier to generalize about the First World than it is about the Third. Nonetheless, if you study Table 1, you will see some patterns in it. This section of the chapter examines these patterns and others.

The dimensions of poverty

Because the table is organized by ranking countries in order of ascending per capita GNP, it is not very surprising to find that per capita output increases as you move down the table. You may, however, be struck by how many people live in countries with very low output. A good way to think about the GNP figures is the following:

1. In 1979, the low-income countries had 58 percent of the population of all countries, but they produced only 6 percent of the combined GNP.
2. The middle-income countries had another 25 percent of the population and produced 17 percent of the GNP.
3. The industrial market economies, with only 17 percent of the population, produced 77 percent of the GNP.

You might also notice from this table that none of the industrial market economies was stagnant during 1960–1979. Even the slowest-growing members of this group had annual increases in per capita

Table 1 Indicators of economic development

	Population (millions) Mid-1979	Area (thousands of square kilometers)	GNP per Capita (dollars) 1979	Average Annual Growth (percent) 1960–1979	Adult Literacy Rate (percent) 1976	Life Expectancy at Birth (years) 1979
Low-income countries	**2,260.2**	**33,778**	**230**	**1.6**	**51**	**57**
China and India	**1,623.7**	**12,885**	**230**	**. . .**	**54**	**59**
Other low-income	**636.5**	**20,893**	**240**	**1.8**	**43**	**50**
1 Kampuchea, Dem.	. . .	181	. . .	. . .	. . .	. . .
2 Lao PDR	3.3	237	. . .	. . .	. . .	42
3 Bhutan	1.3	47	80	—0.1	. . .	44
4 Bangladesh	88.9	144	90	—0.1	26	49
5 Chad	4.4	1,284	110	—1.4	15	41
6 Ethiopia	30.9	1,222	130	1.3	15	40
7 Nepal	14.0	141	130	0.2	19	44
8 Somalia	3.8	638	. . .	—0.5	60	44
9 Mali	6.8	1,240	140	1.1	10	43
10 Burma	32.9	677	160	1.1	67	54
11 Afghanistan	15.5	648	170	0.5	12	41
12 Vietnam	52.9	330	. . .	. . .	87	63
13 Burundi	4.0	28	180	2.1	25	42
14 Upper Volta	5.6	274	180	0.3	. . .	43
15 India	659.2	3,288	190	1.4	36	52
16 Malawi	5.8	118	200	2.9	25	47
17 Rwanda	4.9	26	200	1.5	. .	47
18 Sri Lanka	14.5	66	230	2.2	85	66
19 Benin	3.4	113	250	0.6	. . .	47
20 Mozambique	10.2	783	250	0.1	. . .	47
21 Sierra Leone	3.4	72	250	0.4	. . .	47
22 China	964.5	9,597	260	. . .	66	64
23 Haiti	4.9	28	260	0.3	. . .	53
24 Pakistan	79.7	804	260	2.9	24	52
25 Tanzania	18.0	945	260	2.3	66	52

Table 1 Indicators of economic development—cont'd

	Population (millions) Mid-1979	Area (thousands of square kilometers)	GNP per Capita (dollars) 1979	Average Annual Growth (percent) 1960–1979	Adult Literacy Rate (percent) 1976	Life Expectancy at Birth (years) 1979
26 Zaire	27.5	2,345	260	0.7	15	47
27 Niger	5.2	1,267	270	—1.3	8	43
28 Guinea	5.3	246	280	0.3	20	44
29 Central African Rep.	2.0	623	290	0.7	. . .	44
30 Madagascar	8.5	587	290	—0.4	50	47
31 Uganda	12.8	236	290	—0.2	. . .	54
32 Mauritania	1.6	1,031	320	1.9	17	43
33 Lesotho	1.3	30	340	6.0	52	51
34 Togo	2.4	57	350	3.6	18	47
35 Indonesia	142.9	1,919	370	4.1	62	53
36 Sudan	17.9	2,506	370	0.6	20	47
Middle-income countries	**985.0**	**38,705**	**1,420**	**3.8**	**72**	**61**
Oil exporters	**324.8**	**13,781**	**1,120**	**3.1**	**64**	**57**
Oil importers	**660.2**	**24,924**	**1,550**	**4.1**	**76**	**63**
37 Kenya	15.3	583	380	2.7	45	55
38 Ghana	11.3	239	400	—0.8	. . .	49
39 Yemen Arab Rep.	5.7	195	420	10.9	13	42
40 Senegal	5.5	197	430	—0.2	10	43
41 Angola	6.9	1,247	440	—2.1	. . .	42
42 Zimbabwe	7.1	391	470	0.8	. .	55
43 Egypt	38.9	1,001	480	3.4	44	57
44 Yemen, PDR	1.9	333	480	11.8	27	45
45 Liberia	1.8	111	500	1.6	30	54
46 Zambia	5.6	753	500	0.8	39	49
47 Honduras	3.6	112	530	1.1	60	58
48 Bolivia	5.4	1,099	550	2.2	63	50
49 Cameroon	8.2	475	560	2.5	. .	47
50 Thailand	45.5	514	590	4.6	84	62
51 Philippines	46.7	300	600	2.6	88	62

Table 1 Indicators of economic development—cont'd

	Population (millions) Mid-1979	Area (thousands of square kilometers)	GNP per Capita (dollars) 1979	Average Annual Growth (percent) 1960–1979	Adult Literacy Rate (percent) 1976	Life Expectancy at Birth (years) 1979
52 Congo, People's Rep.	1.5	342	630	0.9	. . .	47
53 Nicaragua	2.6	130	660	1.6	90	56
54 Papua New Guinea	2.9	462	660	2.8	. . .	51
55 El Salvador	4.4	21	670	2.0	62	63
56 Nigeria	82.6	924	670	3.7	. . .	49
57 Peru	17.1	1,285	730	1.7	80	58
58 Morocco	19.5	447	740	2.6	28	56
59 Mongolia	1.6	1,565	780	3.0	. . .	63
60 Albania	2.7	29	840	4.2	. . .	70
61 Dominican Rep.	5.3	49	990	3.4	67	61
62 Colombia	26.1	1,139	1,010	3.0	. . .	63
63 Guatemala	6.8	109	1,020	2.9	. . .	59
64 Syrian Arab Rep.	8.6	185	1,030	4.0	58	65
65 Ivory Coast	8.2	322	1,040	2.4	20	47
66 Ecuador	8.1	284	1,050	4.3	77	61
67 Paraguay	3.0	407	1,070	2.8	84	64
68 Tunisia	6.2	164	1,120	4.8	62	58
69 Korea, Dem. Rep.	17.5	121	1,130	3.5	. .	63
70 Jordan	3.1	98	1,180	5.6	70	61
71 Lebanon	2.7	10	. .	. .	. .	66
72 Jamaica	2.2	11	1,260	1.7	. .	71
73 Turkey	44.2	781	1,330	3.8	60	62
74 Malaysia	13.1	330	1,370	4.0	60	68
75 Panama	1.8	77	1,400	3.1	. .	70
76 Cuba	9.8	115	1,410	4.4	96	72
77 Korea, Rep	37.8	98	1,480	7.1	93	63
78 Algeria	18.2	2,382	1,590	2.4	35	56
79 Mexico	65.5	1,973	1,640	2.7	82	66
80 Chile	10.9	757	1,690	1.2	. .	67
81 South Africa	28.5	1,221	1,720	2.3	. .	61
82 Brazil	116.5	8,512	1,780	4.8	76	63
83 Costa Rica	2.2	51	1,820	3.4	90	70
84 Romania	22.1	238	1,900	9.2	98	71

Table 1 Indicators of economic development—cont'd

	Population (millions) Mid-1979	Area (thousands of square kilometers)	GNP per Capita (dollars) 1979	Average Annual Growth (percent) 1960–1979	Adult Literacy Rate (percent) 1976	Life Expectancy at Birth (years) 1979
85 Uruguay	2.9	176	2,100	0.9	94	71
86 Iran	37.0	1,648	. .	. .	50	54
87 Portugal	9.8	92	2,180	5.5	70	71
88 Argentina	27.3	2,767	2,230	2.4	94	70
89 Yugoslavia	22.1	256	2,430	5.4	85	70
90 Venezuela	14.5	912	3,120	2.7	82	67
91 Trinidad and Tobago	1.2	5	3,390	2.4	95	70
92 Hong Kong	5.0	1	3,760	7.0	90	76
93 Singapore	2.4	1	3,830	7.4	. .	71
94 Greece	9.3	132	3,960	5.9	. .	74
95 Israel	3.8	21	4,150	4.0	. .	72
96 Spain	37.0	505	4,380	4.7	. .	73
Industrial market economies	**671.2**	**30,430**	**9,440**	**4.0**	**99**	**74**
97 Ireland	3.3	70	4,210	3.2	98	73
98 Italy	56.8	301	5,250	3.6	98	73
99 New Zealand	3.2	269	5,930	1.9	99	73
100 United Kingdom	55.9	245	6,320	2.2	99	73
101 Finland	4.8	337	8,160	4.1	100	73
102 Austria	7.5	84	8,630	4.1	99	72
103 Japan	115.7	372	8,810	9.4	99	76
104 Australia	14.3	7,687	9,120	2.8	100	74
105 Canada	23.7	9,976	9,640	3.5	99	74
106 France	53.4	547	9,950	4.0	99	74
107 Netherlands	14.0	41	10,230	3.4	99	75
108 United States	223.6	9,363	10,630	2.4	99	74
109 Norway	4.1	324	10,700	3.5	99	75
110 Belgium	9.8	31	10,920	3.9	99	72
111 Germany, Fed. Rep.	61.2	249	11,730	3.3	99	73
112 Denmark	5.1	43	11,900	3.4	99	75
113 Sweden	8.3	450	11,930	2.4	99	76
114 Switzerland	6.5	41	13,920	2.1	99	75

Source: World Bank, *World Development Report.*

GNP of around 2 percent. But many of the countries in the low-income category in 1979 were just about where they were in 1960, and some were worse off. In the middle-income group, high rates of growth over 1960–1979 were fairly common. Indeed, that is how many of them came to be in the middle-income group in 1979. If Kenya, for example, had had zero growth from 1960–1979, it would have ranked between Rwanda and Sri Lanka in 1979. If Yemen had not grown, it would have been the poorest country listed in the table. Thus, although growth was the rule in the First World, there was no rule for the Third World. Some countries grew and some did not.

Of course, many measures of personal welfare accompany differences in prosperity. Per capita consumption figures roughly parallel the differences in GNP. However, since the more wealthy countries tend to save more than the very poorest countries, the differences in consumption are less pronounced than the differences in GNP.

One particularly important measure of well-being is food consumption. Money value comparisons from country to country are hard to interpret, since people in underdeveloped countries farm and process more food for home consumption than people in industrial countries do. But the World Health Organization estimates that in 1977 the residents of the low-income countries consumed about 98 percent of the daily calorie levels needed to sustain normal activities and health, those in middle-income countries consumed 109 percent, and those in industrial market countries consumed 131 percent. Note that the figure for the poorest countries was less than 100 percent. This was an average across 36 countries; the figures for Chad and Ethiopia were about 75 percent.

Life expectancy figures correlate closely with GNP per capita. As Table 1 shows, newborn children in Japan and Sweden have a life expectancy nearly twice as great as the newborn in Chad and Ethiopia. In the poorest countries, about 1 child in 5 dies in its first year. In the industrial market countries, the figure is about 1 in 100.

Differences in literacy rates also reflect differences in income (and partly cause them). In the poorest countries as a whole, only about half the adults can read and write. In the industrial countries, nearly all adults are literate.

These statistics tell a dreary tale. And since they are based on national averages, they also hide much of the story. Income is distributed within the Third World countries at least as unequally as it is in the First World. If the average Ethiopian gets three fourths of the food calories necessary to live a normal, healthy life, what happens to the poorest Ethiopians?

Industrial structure

Despite the great variety of economic types to be found among the Third World countries, there are structural patterns that distinguish the poorest countries from the middle-income countries, and distinguish both from the industrial market countries. One obvious factor is the degree of urbanization. In 1980, only 17 percent of the population of the low-income countries lived in urban areas. In the middle-income countries, 50 percent lived in urban areas, and in the industrial market countries, 77 percent. (However, the two most urbanized states in the world are Hong Kong and Singapore. They have both had spectacular success in developing their manufacturing industries. Because of this, these two countries have the highest and among the most rapidly growing incomes outside Europe, North America, and Japan.)

Along with the tendency for underdeveloped countries to be rural and developed countries to be urban, there is a cor-

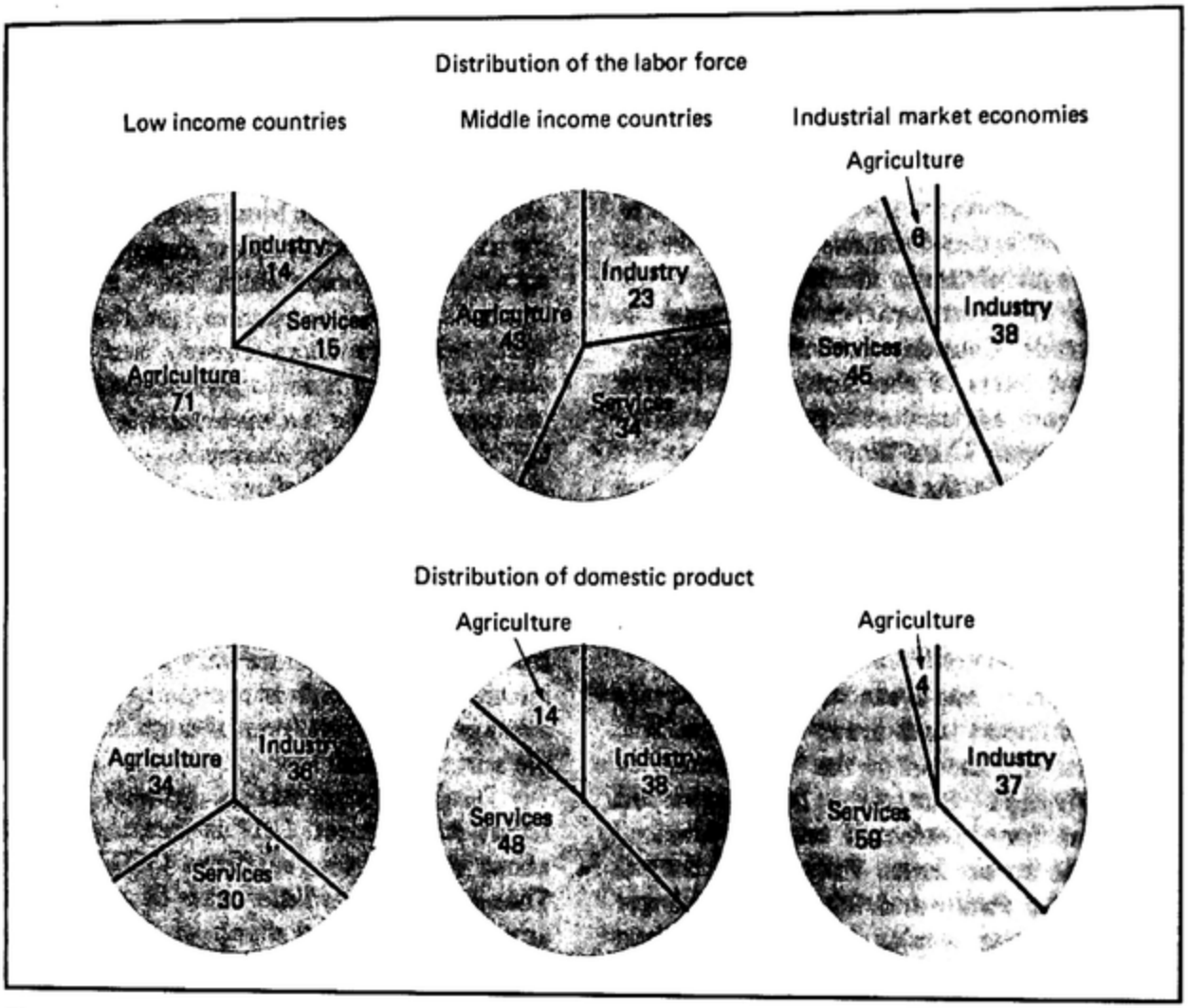


Figure 2 The distribution of the labor force and domestic product among agricultural, industrial, and service sectors 1979 (percent)

Less developed countries are heavily agricultural. More developed countries have relatively larger industrial and service sectors.

Source: World Bank, *World Development Report.*

responding pattern to industrial structure. Whether you look at output or employment, you can see from Figure 2 that less developed countries depend heavily on agriculture, a sector that is not very important in most of the wealthier countries. This in itself would make the Third World poorer than the First World, since throughout the world the product of agricultural labor is less highly valued than that of labor in other sectors.

But there is something else to note from these numbers. In the industrial market economies, the labor force employed in agriculture is only about two-thirds as productive as labor is in the rest of the economy. In the low- and middle-income countries, agricultural labor is only about one-fifth as productive as labor in the other sectors. In less developed countries, labor is far less productive in every sector than it is in the industrialized countries. But agriculture is the most backward of the backward industries. Throughout much of the Third World, agriculture is a form of subsistence production. Peasant families can barely feed themselves on their small farms. Because they produce so

little for market, they are not even helped by rising world agricultural prices.

There is, of course, no reason why an agricultural country cannot be prosperous. The United States did quite well by specializing in agriculture in the 19th century, exploiting its comparative advantage in world trade. Canada, Australia, New Zealand, and parts of the United States still benefit from agricultural specialization today. But much Third World agriculture lies entirely outside the market economy. Peasants produce only for themselves and are the victims of their own low productivity.

Uneven development

Every country has areas that are relatively well developed and areas that are not: Compare Manhattan to the hill country of West Virginia. But some less developed countries have perfected uneven development into an art form. Take Brazil, for example. São Paulo and Rio de Janeiro are cities of great wealth and cosmopolitan culture. Yet, Brazil has one of the most unequal income distributions in the world, with higher infant mortality, lower life expectancy, and a lower literacy rate than those of other countries with comparable levels of GNP. In fact, Brazil is two countries, one rich and one poor. The top 10 percent of the income distribution gets half the income. The remaining 90 percent of the people live on the other half. (The top 10 percent in the United States only gets about 25 percent of the income.)

Brazil is not unique. Mexico is much the same. The bottom 20 percent of the economic ladder there lives on 3 percent of the income, compared to 2 percent in Brazil. Mexico City and Acapulco are gathering places for the jet set. But millions of Mexican peasants are desperate enough to work as illegal farm laborers in Texas or California rather than try to make a living at home.

In its extreme form, ***uneven development*** produces a ***dual economy.*** There is a modern, developed, wealthy sector, usually based on either tourism or a commodity export industry. Then there is a poor agrarian economy. Both lie within the same national boundaries, yet the developed sector is more closely tied to the world economy than to its domestic agrarian partner. This is characteristic of many OPEC countries, for example.

This type of dualism makes per capita GNP figures somewhat misleading as indicators of economic growth. The Middle Eastern oil-producing states have the highest per capita incomes in the world. Yet, most of their people are very poor. A small number of very high incomes can make the average quite high, even though most people's incomes are low.

Dualism also makes growth figures tricky to interpret. GNP in Brazil grew at about 9 percent a year from 1970 to 1979. This is an extraordinary rate of growth by any standard. But much of it seems to have been concentrated in the already highly developed sector of the country. In a sense, it was the growth of an already developed economy. Many of the undeveloped parts of Brazil remained undeveloped.

Why are they underdeveloped?

When the Third World economies used to be called backward, it was obvious the users of the word thought they should be brought up to date. Although modern terminology doesn't carry the same clear implication, the presumption is still there. Development is the norm; underdevelopment is the exception.

This seems presumptuous. Can what is true of three fourths of the world's people be the exception, and what is true of one fourth be the rule? We often think that way. Many of us also think of the tradi-

tional nuclear family—working father, housewife mother, and their children—as the normal living arrangement. We consider other arrangements to be exceptions. Since fewer than 20 percent of American households are like this nowadays, a lot of us must be living exceptional lives.

There is a real sense, however, in which underdevelopment is not a norm. It is a very oppressive condition, objectively destructive of human development. The people caught up in it don't like it. They want out. One of the fundamental problems facing humanity is how to get them out.

In trying to figure out why the Third World is underdeveloped and what might be done about it, it seems helpful to think about how the First and Second Worlds developed. After all, 200 years ago, the United States' per capita GNP was on a par with that of some of the poorer Third World countries today. The Soviet Union of 60 years ago was in a similar situation. Can't we just see how the developed countries got that way and apply the lessons of their history to the Third World?

Yes and no. Yes, because the high-technology, capital intensity, and labor skill that characterize developed countries virtually define development. The Third World cannot be wealthy without being productive, which means using more modern productive methods. No, because the Third World is the product of a very different history from that of the First and Second Worlds, one that involves not only differences in climate, population, culture, and national tradition, but also the fact that there is already a developed part of the world today, as there was not when the First World developed.

In one sense, therefore, the Third World economies are poorer versions of the developed economies. They need modern technology and capital—both human and physical. If they develop, they will have to undergo some of the economic transformations that the American economy went through. In another sense, they are qualitatively different, not just poorer. Both their traditions and their relations with the developed world have blocked their development in the past and may continue to do so.

Capital, saving, and growth

One of the most obvious differences between developed and underdeveloped economies is that the former have a lot more physical capital per worker than the latter. Since machinery is the bearer of modern technology, the lack of capital in the Third World means technological backwardness in production.

The First World got its capital through a drawn-out process of accumulation, at times rapid and at times slow. The Second World (The Soviet Union and its satellites) forced the pace of its development through a crash program of accumulation by state enterprises. In both cases, a substantial share of current output had to be set aside for accumulation. The developed capitalist world did this largely through private saving and investment. The socialist world achieved the same result through state saving and investment.

Some prominent development economists have argued that many of the less developed countries have so little capital that labor is not really scarce. Such countries suffer from ***disguised unemployment.*** Most people are working, but output would not suffer if the work force were reduced. This may be particularly true in agriculture, which tends to be very *labor intensive* in countries such as India and China. Where this is so, urban migration and industrial development could take place without impairing the food supply. It is as though development could take place with essentially *unlimited supplies of labor.*

If the labor supply is not really a binding constraint on development, the arithmetic of saving and growth is particularly simple. With no problem of diminishing returns to capital, output is proportional to the capital stock. Consequently, the rate of growth of output depends only on the average product of capital and the share of output saved:

$$\frac{\Delta Y}{Y} = \frac{\Delta K}{K} = \frac{Y}{K} \times \frac{\Delta K}{Y}$$

Rate of growth of output	Rate of growth of capital	Average product of capital	Investment as a share of output

where Y equals annual net output, K equals capital, ΔY equals the annual change in output, and ΔK equals the annual change in capital, or net investment. The growth rate equals the arithmetic product of the average product of capital (Y/K) and the share of output devoted to net investment ($\Delta K/Y$).

This relationship between saving and growth is illustrated in Figure 3. Each of the two lines coming from the origin represents a different technology. *OA* corresponds to a high capital-labor ratio and, therefore, to a low average product of capital, although a high average product of labor. *OB* corresponds to a low capital-labor ratio, and a high average product of capital, although a low average product of labor. As long as there is surplus labor, the average product of labor isn't important. What matters is the average product of scarce capital. For maximal growth, the surplus-labor economy should be as far out as possible along the line corresponding to the least capital intensive technique—that is, the one that gets the most additional output per unit of capital accumulated.

If the underdeveloped countries are going to reach tolerable levels of per capita output by the end of the century, say, they will have to raise their growth rates a great deal. At a 5 percent growth rate, output doubles every 14 years. At a 10 percent rate, it doubles every 7 years. For the surplus-labor economy, raising the investment rate seems to offer a way to shorten the time needed to achieve a given output target. Therefore, one source of the extreme poverty in these economies is insufficient investment. Except for China, the poorest countries have a lower ratio of investment to output than the faster-growing economies of the middle-income category.

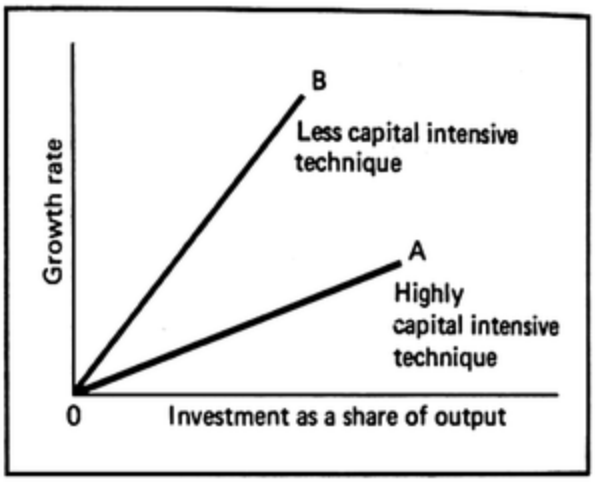


Figure 3 Saving and industrial growth in a surplus-labor economy

In a surplus-labor economy, the growth rate of output is proportional to the share of output invested. If the technology is highly capital intensive, a given saving rate produces less growth than it would if the technology were less capital intensive.

Another reason for the poverty of the less developed countries may be a poor choice of technique. To the extent that these countries choose to develop their industry by employing highly capital intensive, labor-saving techniques borrowed from developed economies, they squander their scarcest resource, capital, and save their surplus resource, labor. It is frequently argued that if Third World countries want to develop, they must create their own ***appropriate or intermediate technology*** rather than borrow from the technological storehouse of the developed

economies. The so-called *advantage of backwardness,* the ability to borrow the latest technology rather than having to develop it, may be an illusion.

Population growth

One way of looking at a country whose population is too large relative to its capital stock is to say that it has too little capital. Another is to say that it has too much population. Which is more correct? For a country like China, whose arable land is densely populated, population is obviously the problem. But in many countries of Africa and South America, much of the good land is empty. These countries are not overpopulated in any absolute sense.

Yet, they have a population problem. It comes from the *growth rate* of the population, not its size. Because their populations grow so rapidly, the capital that they do accumulate does not produce much gain in per capita output.

Between 1970 and 1979, the typical Third World country had a population growth rate of about 2½ percent a year. The First World population grew at about ¾ percent a year. This is a very substantial difference. At these rates, it would take 93 years for the First World population to double. Over the same time span, the Third World population would increase by a factor of 10.

The population growth rate is the difference between two figures, the birth rate and the death rate. These are usually measured in numbers of annual births or deaths per thousand of population, but they can also be expressed as percentages. In 1979, the typical low-income country had a birth rate of about 40 per thousand and a death rate of about 15 per thousand (4 percent and 1½ percent, respectively). The difference, 25 per thousand, produced a population growth rate of 2½ percent a year. In the United States, the corresponding figures were 17 per thousand for births and 9 per thousand for deaths, for a growth rate of about ¾ percent a year.

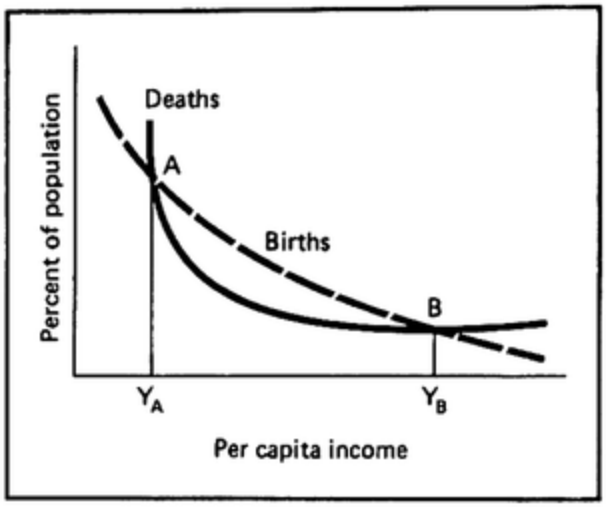


Figure 4 Births, deaths, and population growth in economic development

Both the death rate and the birth rate vary with the degree of economic development. At a low level of per capita income, they are equal and high. At a high level of development, they are equal and low. Thus, population growth is zero (ZPG) at income levels Y_A and Y_B. At any income level in between, the birth rate is higher than the death rate, and population increases over time.

Obviously, a country can have ***zero population growth (ZPG)*** only if its birth and death rates are equal. But they can be equal and high or equal and low. One view of the relationship between population growth and economic development goes like this: At very low levels of development, both the birth and death rates are high. At high levels of development, both are low. In between, there is a ***demographic transition,*** during which the birth rate is much higher than the death rate. This means that a country can have two ZPG populations. One corresponds to a low level of per capita income, the other to a high level. This is illustrated in Figure 4. At Points *A* and *B*, corresponding to per capita income levels Y_A and Y_B, the birth and death rates are equal. But at any intervening income level, the birth rate exceeds

the death rate, and population increases over time.

Comparisons of different countries and studies of the histories of individual countries give some support to this theory, but the data must be interpreted with care. Historical data are particularly tricky, since the technology of health care and family planning is not held constant. Major breakthroughs in public health, such as sprays for the control of disease-carrying insects, were responsible for the sharp decline in death rates in much of the world. These changes reflected the development of scientific and technical knowledge in the world as a whole, not the growth of income in the countries in which they were applied. Moreover, the negative relationship between the birth rate and per capita income is not very strong. Apparently, the decline in the birth rate that takes place during the demographic transition is associated with dimensions of development that go well beyond the level of per capita income. The distribution of the gains from development is particularly important. The drop in the birth rate is most pronounced when much of the population—especially women—sees that the economic opportunities open to it are widening. Women will insist on small families only if they can see clear-cut, attractive alternatives to a lifetime of child rearing.

The theory of the demographic transition is sometimes combined with a theory of capital accumulation to explain why some countries don't grow at all and others grow very rapidly. According to this explanation, a low saving rate at low per capita income levels makes capital grow more slowly than population. At higher income levels, the saving rate is greater and the population growth rate is lower, so that capital grows faster than population. If a country can somehow raise its per capita income level high enough, its saving rate will be high enough to let its capital-labor ratio grow, and it will "take off" into steady development. But if it does not reach the income level at which capital grows faster than population, it remains stuck in a *low-level equilibrium trap.*

This theory has been called too mechanical and simplistic, reducing the complex process of growth to a relationship between two variables. Doubtless, the criticism is well taken. But the tendency for population to grow very rapidly in early stages of development does erect a major barrier to raising the capital-labor ratio.

Education and development

Careful historical studies of First World countries show a clear relationship between the spread of education throughout the population and the spread of economic development. In part, the influence of education on growth comes from the accumulation of human capital. The more educated the population is, the more its skills are directly applicable to the process of production, and the more quickly it can assimilate new skills. But historians think there is more to it than that. Wide-scale development requires the innovative participation of much of the population. Education frees people from the inertia of accepted tradition. Through books, newspapers, and magazines, a literate population becomes aware of alternative ways of doing things. This makes it receptive to novelty and innovation. This receptivity is part of a "modern" attitude that is a precondition for participating in development.

Table 2 compares the extent of education in low-income, middle-income, and industrial market countries. As you can see, the low- and middle-income countries are particularly far behind in secondary and higher education. These are the edu-

Table 2 ***Primary, secondary, and higher education enrollments as a percent of population age group, compared to adult literacy rate, 1978***

	Primary	Secondary	Higher	Literacy Rate
Low-income countries	83	36	3	51
Middle-income countries	95	41	11	72
Industrial market countries	100	89	37	99

Source: World Bank, *World Development Report.*

cational levels that make the greatest contribution to widening people's horizons.

Agriculture and development

A country caught in a low-level equilibrium trap has a weak tendency to accumulate capital and a strong tendency for its population to grow. Any spurt of capital accumulation is swamped by the population response. These strong and weak tendencies seem to describe some aspects of the situation underdeveloped countries are in. But, in turn, they themselves need to be explained.

Part of the explanation lies in the social structure of Third World countries. Most of the population in the poorest countries—over 80 percent—lives in rural areas. The typical production unit is the *extended peasant family,* consisting of several generations housed together and working a small plot of land in common. Typically, they supply most of their own needs through subsistence farming and handicrafts. They also produce one or more cash crops on a small scale and supply some of their needs out of the proceeds from selling them. Both the ***subsistence and cash crop agriculture*** are conducted by age-old methods and are very unproductive. The result is a very meager standard of living.

This social and economic structure is almost uniquely unsuited to rural development. First of all, the sheer poverty of subsistence agriculture makes it very hard for peasant farmers to save, buy machinery, fertilizer, and seed, and improve their productivity. Each little farm is caught in its own low-level trap. Second, the precariousness of life makes the countryside resistant to change. Every family that pursues traditional farming methods knows that it can eke out a living that way, except in times of drought or flood. If it tries something different, and especially if it abandons food production in favor of specialization in a cash crop, who knows what will happen? Ordinary prudence dictates that the farm not depart very far from the traditions that have kept it intact. Third, the individual members of an extended family are inhibited from "going it alone" and trying something different. Partly this results from social pressure within the family, partly from a lack of individual control over resources, and partly from the obligation to share the fruits of success and the costs of failure with the whole family. Fourth, raising a large family is a way to ensure both the continuation of the productive unit despite the high death rate and support for one's old age. Thus, the peasant farm is incapable of capital accumulation, resistant to technical change, and conducive to population growth. Wherever it is the dominant social and economic form, it is a formidable barrier to development. And the barrier is not the result of the stupidity or irrationality

of peasant farmers. They are doing the best they can, given their circumstances.

In some countries, agricultural land is the property of landlords who own large tracts that they let out to the peasants, either for a fixed rent or for a share of the crop. This mode of economic organization produces an income for the landlord that might be a source of capital accumulation. Indeed, landlords often develop their property in ways that are almost impossible for small peasant landholders—especially through the construction of large irrigation facilities. However, they rarely divert their saving to finance industrial development, which is a threat to their political control.

Much the same is true for plantation agriculture, which is large-scale farming using landless wage labor. Again, there is a source of saving that might be used for capital formation. Many plantations are foreign owned, however. Their owners have no particular incentive to invest their profits in the country in which they originate. Native plantation owners also may have no particular incentive to invest at home except to develop their own plantations.

In any case, large-scale landholding rarely provides a source of capital accumulation for industrial development. The reason is that in the Third World nearly every government that comes to power committed to economic development has a peasant political base. Its political appeal to the peasantry usually stems from its commitment to *land reform*. And most land reform programs entail breaking up large landholdings and distributing the land to peasant families. This tends to strengthen small-scale peasant agriculture, the very form of social and economic organization that is such a barrier to development.

Relationships with the First World

Most of the Third World was once part of the colonial system controlled by Western Europe and the United States. When the major First World countries were becoming industrialized, their colonies supplied them with food and raw materials. The colonial powers rarely encouraged development in their dependencies unless it complemented their own industrial needs. This meant investing in mines and plantations, but not in manufacturing. Often, it also meant erecting a system of tariffs and colonial preferences that encouraged trade between the colonies and their colonial rulers, but discouraged trade between the colonies and other parts of the world. Even when there was local capital accumulation, as in the American colonies before the Revolutionary War, the tariff system effectively discouraged industrial development.

In itself, this form of complementarity did not prevent the colonial areas from becoming relatively prosperous. Specialization in forestry, agriculture, mining, and petroleum extraction is not necessarily a one-way ticket to the poorhouse. But four factors made it work out that way in many colonies. First, many ***primary products*** (that is, products of forests, farms, and mines) had a rather low demand elasticity with respect to world income. This meant that colonial industries producing such products did not develop as fast as the rest of the world. Second, the mines and plantations were often owned by residents of the imperial power. Profits were funneled to the owners rather than invested in the colonies. Third, the colonial relationship encouraged the development of a dual economy with a modern sector linked closely to the colonial power, and a backward subsistence sector linked neither to the imperial country nor to its developed domestic counterpart. Fourth, the colonial

governments had no interest in educating the subject populations beyond what was necessary to enable them to work in industries that were useful to the colonial powers.

The end of colonialism after World War II did not change the economic patterns that grew up under it. Political independence did not bring economic independence. The former colonies were still heavily committed to the export of primary products because these were the only developed industries they inherited from their colonial past. Often, they continued to trade with the very countries that had recently owned them. Moreover, many plantations and mines remained the property of their previous colonial owners.

You might suppose that the hangover of colonialism would pass after a few decades, with no special remedy needed to make it go away. If this were true, the sharp lines of specialization that separated the former colonies from the First World would gradually be replaced by lines based on population density, climate, and natural resource endowments.

In fact, the old lines of specialization do not disappear easily because of the importance of labor skills and capital in determining the pattern of comparative advantage. The countries that industrialized early created for themselves a comparative advantage in industrial production by accumulating capital and educating their populations. The other side of the process was the creation in the rest of the world of a comparative advantage in primary production. This pattern of advantage poses a dilemma for the Third World countries. If they want to industrialize, they must work at cross-purposes to the pattern of comparative advantage. As you know, this can be done only at the cost of giving up the gains from specialization and trade.

How can they develop?

All underdeveloped countries have *development strategies.* A development strategy is simply a political policy toward the economic problem of underdevelopment. Even total reliance on individual enterprise is an implicit strategy, although no contemporary nation pursues it. All Third World countries rely on some form of government policy that explicitly promotes development. In doing so, they are only taking pages from the history books of the First and Second Worlds.

Capitalism and socialism

A major decision about development strategy centers on whether the country will develop with capitalist or socialist institutions. Such decisions are sometimes made at the polls, but most often on the barricades or the battlefield. Korea was divided into the socialist North and the capitalist South after World War II, and the division was confirmed by the Korean War. Chile moved toward socialism after the elections of 1970 but back to capitalism after the coup of 1973.

The most generally accepted distinction between the capitalist and socialist systems of economic organization focuses on who owns and controls the means of production—the factories, shops, farms, machinery, highways, and harbors. Under capitalism, they are privately owned and controlled by managers responsible to the owners. Under socialism, they are publicly owned and controlled by managers responsible to the government.

This distinction does not sort the world's countries into two camps very precisely, since all countries have some elements of both socialism and capitalism. In the Soviet Union, for example, privately controlled farms are quite important. In

socialist Yugoslavia, the individual enterprises have a good deal of autonomy in making production and investment decisions. In the United States, utilities are often publicly owned and highways almost always are. In capitalist Italy, Fiat is a private corporation, but Alfa Romeo is government owned.

The governments of countries in which large-scale industry is state owned have very direct control over resource allocation in the industrial sector. But don't imagine that countries with capitalist industry can't control their industrial development. Tariffs, quotas, export licenses, foreign exchange restrictions, taxes, subsidies, and credit rationing can tailor a straitjacket of control that is every bit as confining as state ownership. The specific policies that make up the development strategy and the vigor with which they are pursued are as important as who owns the means of production.

Capitalism can mobilize individual aspirations and energies on a far wider scale than socialism has been able to. In the right cultural setting, it is unsurpassed as an organizational form for transforming economy and society.

You might suppose, however, that socialism would have inherent advantages over capitalism in allocating resources, if not in efficiency, responsiveness to individual needs, and ability to mobilize individual initiative. In some respects, this is true. Socialist countries need not devote resources to the production of frivolities just because some of their residents want to and can pay for them. But in any very poor country, resource allocation is circumscribed by the desperate needs of the population for food, shelter, clothing, and health care. Decisions to use resources for other purposes entail great human hardship, whether the country is capitalist or socialist. All humanitarian governments are constrained by this fact. And callous governments can be found in both capitalist and socialist camps. The Soviet Union killed off millions of peasants when it industrialized in the 1930s. Great Britain's government permitted its capitalists and landlords to drive millions off the land to emigrate to the cities in the 18th and 19th centuries.

Whether capitalist or socialist, the government of a developing country needs to be stable. Many investments in both physical and human capital have their effects over a very long period. Neither capitalist nor socialist planners will actively pursue long-term investments unless they expect to be around to enjoy the results. Extreme political instability makes people preoccupied with immediate results and discourages their faith in the distant future.

Capital accumulation

As you know, a shortage of capital and an accompanying low level of productivity are the hallmarks of underdevelopment. Every development strategy must face the problem of capital accumulation. Although the share of output devoted to capital formation is not a perfect predictor of the rate of growth, raising the rate of investment is one of the most reliable ways to grow faster.

Unfortunately, knowing that this is true does not immediately produce a development plan. Too many questions have to be answered before a specific policy can be formulated.

> *Where will the resources come from?* In a country that is fully employed, an increase in the share of output devoted to investment has to come at the expense of something else. Personal consumption, public consumption, or exports have to be cut, or imports have

to be increased. If there is open or hidden unemployment, however, the choice is not so hard. Countries like China have successfully raised their investment by mobilizing existing resources more effectively.

How will the investment be financed? Most underdeveloped countries have poorly developed financial intermediaries and capital markets, partly because domestic saving is meager. The main means of financing higher investment are increased taxation and public spending, or greater dependence on foreign aid and lending, both public and private.

Who is going to undertake the investment? Where domestic entrepreneurship is well developed, or foreign capitalists are interested, this is not a problem. But in much of the world, entrepreneurship must come from the government. And where the investment is to be done by domestic or foreign capitalists, some financial incentives must be provided by government policy.

What industries will be developed? Any program of development requires some investment in social overhead capital—the transportation and public utility sectors. Beyond that, there are many options. Agriculture can be favored, with a fairly immediate payoff for the standard of living. Industrial development has a slower payoff, but it may offer more hope for self-sustaining growth in the long run. Within industrial development, emphasis can be put on sectors that reduce dependence on imports or promote exports, a matter that will be discussed in a few pages. Investment can be spread across many industries to ensure balanced growth in complementary sectors of the economy. Or it can be deliberately unbalanced to create tensions that will induce entrepreneurship that fills the gaps.

What kinds of technology will be invested in? If the investment goods are imported from the First or Second World, they will embody the technology favored by the exporting countries' capital goods industries. But if they are produced domestically, they can embody intermediate technology. This means that they will fit into a relatively labor intensive setting and will not require labor skills that are very scarce in underdeveloped countries.

As you can see, a country whose planners intend to embark on a program of intensified capital accumulation has many options. There are nearly as many sensible-sounding possibilities as there are countries to try them out. And there seems to be no tried and true recipe for success that works everywhere.

Population control

If an underdeveloped country wants to embark on a big push to increase its capital-labor ratio, it is more likely to succeed if it can control its rate of population growth. With sufficient economic development, the population problem would take care of itself, much as it has in the First World. But a spontaneous drop in the birth rate won't happen unless people want smaller families. Raising the national growth rate of per capita income is just not a persuasive reason to keep people from having children.

Economic planners who see the benefits of lower population growth usually promote *family planning* as part of an overall development strategy. In the mid-1970s, it cost between $5 and $20 per year to prevent a birth by using conventional

contraceptive devices. On the basis of such figures, increased investment in family planning seems to promise enormous returns from reducing the share of output devoted to child rearing and education. Of course, the returns don't materialize unless people want to have fewer children. Hence, despite birth-control programs in nearly every underdeveloped country, the gains have been small so far—a few tenths of a percent reduction in the birth rate.

Agriculture

The development of agriculture is the crucial element in an overall development strategy. No genuinely underdeveloped country outside OPEC has a sufficiently productive export industry to permit it to neglect agriculture.

The peasant agricultural sector in most countries is the major stumbling block of development policy. In the 1960s and 1970s, there was little growth in per capita food production in the Third World. Growth was mainly confined to large-scale farming, where labor productivity was already 5 to 20 times that in peasant agriculture. Large farms that are integrated into the market economy do not pose problems any different in kind from those of industrial development. They respond to capital accumulation and technical change.

The problem of agriculture in a poor country is really two problems in one. The first centers on the peasants themselves. Because they are poor and malnourished, they desperately need a higher standard of food consumption. The second is the problem of the urban work force, which is similarly poor and malnourished. It too needs more food. Somehow agricultural productivity must be raised sufficiently to resolve the needs of both city and town.

A successful program of agricultural development must have three elements:

1. The technology of farming must be transformed from primitive to modern. There are few places in which more intense application of traditional methods will pay off. Something different is needed.
2. Farmers must stop producing for subsistence and start producing for market. This will make scale economies possible even on relatively small farms, as they specialize in a few cash crops. It will also divert some of the fruits of higher farm productivity to the urban work force.
3. The average farm size will have to be increased to take advantage of scale economies and, especially, of production techniques that are suitable only to large-scale farming.

In any country whose agriculture is now dominated by subsistence farms, progress will necessarily be slow because a *social* transformation will be needed to move toward large-scale farming. Many countries are groping toward alternatives to the family farm—some toward capitalist plantation agriculture and some toward collectivist solutions, like the Chinese communes. But meanwhile, something must be done within the framework of peasant agriculture itself. One step is to increase the security of land tenure for peasants who do not own their land. This will provide them with a direct incentive to preserve and develop the land's productivity. A second step is to develop institutions for sharing risk. Hungry peasants are understandably reluctant to specialize in cash crops, since they thereby commit the survival of their families to the uncertainties of the market. A third step is to develop technology appropriate to small farms that could not use machinery even if they had it. Innovations must be tailored to fit the social structure of peasant farming, and not the other way around.

A good instance of a major breakthrough in agricultural technology that

has *hurt* peasant farming more than it has helped is the so-called *green revolution.* This has been the development during the last 30 years of extremely high yield varieties of corn, wheat, and rice that have more than doubled the grain output of land and labor. The main problem with the new varieties is that they require *complementary inputs* of irrigation water and chemical fertilizers, pesticides, and herbicides. This makes them suited only to plantation agriculture, and useless for small-scale farming. Another related problem is that the hybrid seeds and complementary inputs are only available from the *agribusiness* corporations of the First World. Any country that wants to join the green revolution must have sufficient foreign exchange earnings to permit it to import seeds and chemical inputs.

Ironically, the green revolution may actually have retarded the development of peasant agriculture and harmed the most progressive peasant farmers. To the extent that the new techniques were adopted on large farms, they increased grain supplies and depressed prices. Small farmers who produced grain for market using traditional techniques were impoverished. Their fears about the uncertainties of cash crop farming were confirmed.

Trade and development

As you know, the case for free trade rests on the argument that world income is higher if countries specialize than it is if they diversify and try to be self-sufficient. This seems to suggest that it is better for some countries to specialize in primary products and others in manufactures.

But comparative advantage is an acquired characteristic. England has a comparative advantage in manufacturing and New Zealand in sheep raising precisely because England has been industrialized and New Zealand has not. If the capital stock of England and the skills of its work force were magically transported to the opposite side of the globe, the grasses of England permitted to grow, and the sheep transported to England, the pattern of comparative advantage would be reversed, since it is based on history, not on differences in natural resources.

In a sense, the Third World has a comparative advantage in being poor. The immediate interests of these countries lie in continuing to produce primary products for the world market. Any substantial diversion of resources from the present export industries will lower current output. A development program that reallocates resources away from the production of primary products has to be justified on long-run grounds.

Much of the long-run justification rests on the undesirability of being in the primary products business. It is undesirable first because the world income elasticity of demand for foodstuffs and raw materials is quite low. Any countries specializing in their production are committed to producing for a slowly growing market. Less developed countries have, in fact, been losing their share of the world export market over the last three decades. If their export industries grow faster than demand, they are committed to a long-term decline in the terms of trade, since the prices of their exports will fall if supply outstrips demand. Second, even in the short run, the prices of primary products are unstable. Supplies are affected by weather and fluctuate sharply from year to year. Since demands are very inelastic, fluctuations in price are far greater than fluctuations in quantity. This makes export earnings very unpredictable. Since capital goods must be imported and paid for in foreign currencies, the instability of export earnings constantly disrupts the process of development. A country that is highly dependent on primary products lacks control over its own future. The need to get out of these

short- and long-run difficulties of primary-product markets is one of the major reasons that less developed countries want to diversify their economies. The other is that industrial diversification brings with it cultural diversification and development.

Third World countries have pursued two trade-related strategies of diversification. One is known as ***import substitution,*** a policy of developing domestic sources of supply for goods previously imported. This is an *inward-looking policy*. Falling export earnings are countered by reducing dependence on imports and moving toward self-sufficiency.

The means employed to achieve import substitution are the conventional methods of developing *infant industries*. Tariffs and quotas keep out foreign goods, and tax concessions and other subsidies encourage domestic substitute industries. In practice, import substitution has not worked very well in the countries in which it has been tried extensively, mainly in South America. This seems largely to have been the result of poor planning, and not of an inherent defect in the policy. Countries that have chosen to develop substitutes for motor vehicle imports have been especially unsuccessful. They encouraged foreign firms to build plants behind tariff barriers to produce the same kinds of vehicles the firms had been exporting to the Third World. The foreign firms brought in their own managers and technicians and chose to produce engines and other highly machined parts at home. Only assembly was concentrated inside the less developed countries. Although assembly plants do employ people, if the parts that are assembled are imported, not many benefits spill over to other Third World industries. Not even much foreign exchange is saved.

A more promising strategy is ***export promotion.*** This is an *outward-looking policy* to develop alternative export industries. Taiwan, South Korea, Hong Kong, and Singapore have all had spectacular success in developing light manufacturing industries—shoes, clothing, and electronics. These industries are relatively labor intensive and don't require much highly trained labor or complex capital. Thus, they play to strength, following comparative advantage. The virtue of such a policy is that if the right industries are chosen, the subsidy needed to develop them is small and the payoff quick, so that the cost of resource reallocation is not very great.

The main stumbling block in the path of export promotion is First World protectionism. If Third World countries have a comparative advantage in light manufacturing, First World countries do not. Export promotion in the Third World threatens First World industries, and their governments respond to protect them with tariffs and quotas. The ultimate success of export promotion depends on the willingness of First World governments to permit a restructuring of their own domestic industry.

Foreign capital

Developing countries are subject to scarcity constraints that bind very tightly. One is the limitation on resources to devote to investment. One source of investable resources is domestic saving—the difference between GNP and consumption, both public and private. To increase the flow of saving relative to GNP, however, a developing country must be willing to cut back its consumption. This is very painful in a poor country.

Another source of investable resources is foreign trade. A country that somehow manages to import more than it exports can invest more than it saves. The only way to do this is to get finance capital from abroad.

The ***foreign capital*** that flows into Third World countries comes from three

sources. First, there is a major flow from the First World. Second, there is a much, much smaller flow from the Soviet Union and China to their respective Third World allies. Finally, in the past decade, there has been a substantial and growing flow of finance capital from the richest OPEC countries, which are undeveloped but have enormous current-account surpluses from their oil revenues.

Capital flows to the Third World take four different forms. First, there are outright government grants, which you may think of as *foreign aid*. Second, there are government loans on favorable terms, sometimes called *concessional loans*. Third, there are ordinary *private loans*, no different from loans from one part of the developed world to another. Fourth, there are *direct investments*, in which First World firms build and pay for productive facilities in the Third World.

Table 3 gives some data on capital flows from the First to the Third World in 1980. As you can see, the United States supplied only about 20 percent of such capital in 1980, though it supplied nearly 50 percent in 1975. Relative to GNP, the amounts supplied by the United States are far below those supplied by other industrialized countries.

Foreign government grants and concessional loans are *transfers* from developed countries to underdeveloped countries. They provide more capital than Third World countries could get by relying on private sources. For the countries making the transfers, they represent a deliberate giveaway. While some transfers (like disaster relief) are made for humanitarian reasons, with no ulterior motives, most aid is given to further the interests of the countries making the grants. Military aid (which is not included in Table 3) is a favorite form. Much economic aid is also linked to the formation of military and political alliances.

The granting of military aid and the linking of economic aid to an alliance is not damaging in itself. Free warplanes do a country no economic harm. They may even help if the country would otherwise have used export earnings to buy them. Even if it receives development grants in return for threatening to use those warplanes on its neighbors, the grants can nevertheless buy capital goods. What hurts is the damage to national sovereignty, one of the dimensions of national autonomy. A country that accepts aid with political strings attached has given away control over its own future.

Table 3 Capital flows from First to Third World, 1980

	U.S.A.	Other First World Countries
Amount ($ billions)	13.9	61.2
Percent of GNP	0.5	1.3
Distribution (percent)		
Official development assistance	51	32
Grants	28	23
Concessional loans	23	9
Private loans and investments	31	59
Direct investment	24	9
Portfolio investment and loans	7	50
Other official and private	18	9

Source: Organization for Economic Cooperation and Development, *Development Cooperation*.

The strings attached to economic aid are not always linked to international politics. The United States has frequently chosen to give *tied aid,* which can only be used to buy goods in this country. This commits the recipient to import capital goods embodying American technology and requiring American replacement parts. This kind of aid is not always very helpful. The Soviet Union ties its aid even more tightly, demanding that it have substantial control over the recipient's development strategy. The issue of national autonomy was at the center of the rupture of relations between China and the Soviet Union in the 1950s. The Chinese gave up substantial assistance to preserve their autonomy.

The reluctance of Third World governments to take orders from Washington or Moscow makes them favor *multilateral aid,* coming through international organizations, rather than direct bilateral aid. Some African countries won't even accept technical assistance unless it comes through the United Nations.

It is also apparent that countries that encourage direct investment by foreign firms are inviting interference in their domestic politics. Small wonder that some "nonaligned" countries—Burma, for example—have decided to go it alone, without foreign capital. They trade with the First and Second Worlds, but they avoid any more binding entanglements. It may slow their growth, but they think it is worth it.

Summary

This chapter has been a brief introduction to economic development. Its main points are the following:

1. About half the people in the world in 1980 lived in countries whose annual GNP was below $400 per capita. They were poorly educated, malnourished, and short lived.
2. Underdeveloped countries are economically dominated by unproductive peasant agriculture. The more developed parts of these countries are weakly integrated with the poorer parts.
3. Low income results from low productivity, which, in turn, is the result of a shortage of physical and human capital. Poor countries have trouble accumulating capital, but their population increases rapidly.
4. In large part, the low level of development in these countries is a heritage of their past colonial status and the fact that other countries developed earlier.
5. The choice between capitalist and socialist development strategies is important, but the concrete content of programs is what matters most.
6. Attempts to raise the rate of capital accumulation run into a host of problems with no agreed upon solutions.
7. Population planning offers a way to raise the capital-labor ratio by restricting labor force growth. However, success depends on a sufficient level of development so that people will want to have smaller families.
8. The key to successful development is raising the productivity of agriculture, both to improve the lives of the agricultural population and to create a surplus to feed the urban population. Higher agricultural productivity requires larger farms, production for market rather than subsistence, and modern agricultural techniques.
9. A source of slow growth in Third World countries is their specialization in the production of primary com-

modities, for which world demand grows slowly.

10. Foreign finance capital seems like a promising way to get more physical capital without having to cut consumption but political strings may make it less desirable than domestic sources.

Key concepts

First, Second, and Third Worlds
Uneven development, dual economy
Disguised unemployment
Appropriate or intermediate technology
Zero population growth (ZPG)
Demographic transition
Subsistence and cash crop agriculture
Primary products
Import substitution
Export promotion
Foreign capital

Questions for review

1. Underdeveloped countries are heavily specialized in agriculture, and their agricultural sectors are, on average, very unproductive. How do you reconcile this with the theory of comparative advantage?
2. What is a dual economy? Why is dualism encouraged by direct foreign investment?
3. Explain why specialization in the production of primary products is limiting to growth.
4. Why should underdeveloped countries choose techniques of production that have low labor productivity but high capital productivity?
5. Explain the demographic transition.
6. Why do economists usually favor export promotion rather than import substitution?

37

Marxism and Capitalism

As you read and study this chapter, you will learn:

- how Marxists analyze capitalist production and the society based upon it
- why they think it is exploitative
- why they think it contains contradictions that will cause its own destruction

You must have seen the movie *The Wizard of Oz* at some time or another. Remember how it starts off on a poor farm in Kansas? That part is filmed in black and white. But then, Dorothy and Toto get carried off by a tornado to a strange land called Oz, where everything is brightly colored, unexpected, charming, and sometimes terrifying. In the end, the ruby slippers carry Dorothy and Toto back to the black-and-white world of Kansas as she chants, "There's no place like home," over and over.

Are you sure there's no place like home? The film makers seemed to think there's no place like Oz. They *did* film Oz in Technicolor, and Kansas in black and white.

This chapter invites you to take a journey to another strange land, the analytical perspective of Marxism, from which you can look back at your own society. Viewed from this perspective, the U.S. economy, including Kansas, looks quite different from how it probably looks to you.

Many people grow up as Marxists, not only in the socialist

countries, but also in Western Europe, Japan, and even the United States. Others are converts to Marxism, although they were raised in another intellectual tradition. When Marxists think about capitalist society, they often use different terms and concepts from those used by conventional economists. They dismiss much of economics as *vulgar economy*—an account of the superficial details of economic life that rarely penetrate below the surface seen by the economic actors themselves. They might label some parts of this book as *apologetic*—as the self-justifying beliefs of men and women who live comfortably under capitalism.

We are all caught up in *ideologies*, our systems of belief about how social life fits together at a basic level and about how we relate to it. There is a saying that knowing your own ideology is as hard as smelling your own breath. Yet, it is important to try, since it would be tragic to live out your entire life in the grip of a set of beliefs that you have never subjected to searching criticism. One of the names for the ideology that is common to most Americans is the ***conventional wisdom,*** a term coined by John Kenneth Galbraith. Those who are comfortable with the conventional wisdom find Marxist beliefs puzzling, eccentric, and even dangerous. Marxists, on the other hand, see the conventional wisdom as a caricature of reality. Between the two groups lies an ideological Berlin Wall. No embassy can give you a pass through its checkpoints.

Some people, however, cross this wall on the winds of a tornado. They are ideological converts, who study, ponder, observe, work, struggle, and are fundamentally changed. Afterward, they can't believe they ever lived in Kansas and accepted the conventional wisdom. This will probably never happen to you. But if you study this chapter thoughtfully, with an open mind, you might just learn to look at your own society a little more critically. And if you learn to examine both your own beliefs and the alternatives, you will have grown more free.

To put Marxism in the clearest possible light, this chapter presents it as it might be written by a Marxist. It is mostly sympathetic to Marxist economics, in the same way that the other chapters are sympathetic to the conventional wisdom. However, don't expect the chapter to be propaganda. Marxism is a closely reasoned body of economic and social theory, every bit as challenging as the economics of the conventional wisdom. Marxism is not just the body of thought that guides political and economic practice in the Soviet Union, China, Cuba, and other Marxist countries. It is also an alternative analysis of how capitalism works.

One thing you should remember, though, is that Marxists accept much of conventional economic analysis. When they talk about the surface aspects of economic life, they use concepts like *competition, equilibrium, cost, price, the propensity to consume, and the multiplier.* It is only when they talk about the deep motive forces of history that their vocabulary and way of thinking are distinctly different from what you are used to. This chapter presents what is distinctive, but it presumes much of what you already believe.

Work, value, and surplus value

For most people, work dominates life. Indeed, most of us define ourselves by our jobs. Ask a retired person what she or he "used to do." The answer you will get is, "I was a bricklayer," or "I taught school." Hardly anyone will say, "I used to consume peas and carrots," or "I spent much of my time staring off into space."

The analysis of how work effort is organized is central to Marxist economics. To understand Marxist criticisms of capitalism, then, it is important for you to understand how a Marxist views the links among workers, their work, the products of that work, and the capitalist employers whose property these products become.

Commodities

According to Marxists, at the heart of capitalism is something very commonplace and everyday—the ***commodity.*** You probably think of a commodity as an object that is useful to people. To a Marxist, though, a commodity is something more. It must also be produced for market. Production for market is a very old form of human activity and in itself is not capitalism. But when production for market is combined with the employment of wage labor, it becomes the ***capitalist mode of production.*** Under the capitalist mode of production, the products of wage labor become the property of a capitalist enterprise, or *capitalist*, for short. The capitalist enterprise not only pays wages to its workers but also organizes their work effort, equips them with buildings, machinery, and raw materials, and then markets their products. From the capitalist's viewpoint, the goal of the production effort is to make a profit, and the commodity is only something that is sold to get the profit. From the worker's viewpoint, the goal of the production effort is to earn the wage, and the commodity is only something that is made to earn a wage. Thus, the commodity itself is an orphan that neither the capitalist nor the worker wants for its own sake.

Since the worker incorporates his or her labor into the commodity, when the capitalist appropriates the product, he indirectly also appropriates the labor of the worker. When he sells the product, he conveys this labor to the buyer. You can see that under capitalism, the commodity is the means by which the work effort of some is made available to others. The sale and purchase of commodities are, in effect, the sale and purchase of labor.

Direct and Indirect labor

The quantity effort needed to produce a commodity, measured in hours of labor, is known in Marxist terminology as the ***value*** of the commodity. This value has two parts. First, there is the *direct labor* involved in producing and distributing the product. But there is also *indirect labor,* which is embodied in the material inputs. For example, if a firm produces cotton shirts, there is direct labor involved in sewing the shirts. There is also indirect labor, though, since other workers' labor planted and harvested the cotton, wove it into cotton textiles, and made the sewing machines. Thus, to measure the value of material commodities by the labor they embody, you must include both direct and indirect labor. Of course, if you trace the production of a commodity back through the input-output structure, you find that what was indirect labor at one stage of production was direct labor at an earlier stage. By tracing all commodities back to their source, you find that the value of society's total output of commodities just equals the total amount of labor that went into their production.

From a capitalist's viewpoint, of course, material inputs are simply a major element of cost, and not embodied labor. A capitalist would say that the money or ***capital*** that he has invested in a product has two parts: the expenditures on material inputs (raw materials, machinery, buildings) and the wages paid to direct labor. So, the capitalist firm lays out its money capital to acquire material inputs and direct labor, incorporates the inputs

and labor into a product or commodity, and sells the product to recover its money capital, along with a profit.

Labor, labor power, and surplus value

But where does this profit come from? After all, if the value of commodities is just equal to the labor necessary to produce them, there shouldn't be any room for profits. Isn't the value of the product really more than the labor embodied in it? Doesn't property income in the forms of profit, interest, and rent simply represent an additional contribution that capital and land make to the value of a good? A Marxist would answer these questions with a resounding, "No!"

According to conventional economic theory, both workers and owners of capital get income in proportion to their marginal contributions to production. Workers receive wages equal to the marginal productivity of their labor. Owners of property receive income (rent, interest, and profit) equal to the marginal productivity of the land and capital they own. In this view, since labor, land, and capital all contribute to output, they are all productive inputs. It is at this point that Marxist analysis cuts through vulgar economy and analyzes things at a deeper level. The capitalist sees only that he or she makes an income in proportion to the land and capital he or she owns. Indeed, this income seems to arise from the product of that land and capital. But to a Marxist, it is ridiculous to suppose that land and capital are productive. Do land and capital get up in the morning, put on their overalls, pack their lunch pails, and go to work? Do they ache and sweat? Do they plan and execute their work? No, only people work. Of course, land and capital help people to work more effectively. After all, workers equipped with fertile land, good tools, and high-quality materials will produce more output than workers who are ill-equipped. But to make an analogy, land and capital themselves do not produce things, any more than a church prays to God. Prayers said in church may be more effective than prayers said out of doors, but it is the faithful, not the building, that do the praying.

Where, then, does property income come from? All material commodities are the product of labor. The value of a commodity equals the labor effort embodied in it. Yet, some part of the value of a commodity is appropriated by capitalists, in the form of profit, interest, and rent. Capitalists must, therefore, receive part of the value that labor imparts to the product. This can only happen if workers are receiving wages of less value than their product—if they are *exploited.*

How could this happen? After all, under capitalism, transactions take place in markets, under free contract. Since no one is forced to accept another's offer, exchange should only occur under conditions that benefit both sides. It is easy to see why capitalists would want to enter into an exchange in which they can pay wages of lower value than the labor that they will receive in return. But what about the workers? Why are they willing to exchange more for less? If they were slaves, working under compulsion of the whip, they would have no choice. But it is not easy to see why free labor is willing to enter into an unequal exchange.

One of Marx's major contributions to economic analysis was his resolution of this puzzle. This resolution hinges on the distinction between ***labor*** and ***labor power.***

The value of material commodities, you will remember, is determined by how much labor it takes to produce them. But how, you may ask, is the value of labor determined? Marx himself said that the value of labor is a "blue logarithm." He

used this incongruous phrase to point out that the "value of labor" is itself a nonsense concept. Labor *is* the substance of value. It cannot itself have value, any more than usefulness can have utility.

What the worker sells on the market is *labor power*, which is the *ability and willingness to work. Labor* is the actual use of this power in the workplace. It is not something that is bought and sold. The right question to ask is not what determines the value of labor, but what determines the value of labor power. And the answer to it is straightforward: Labor power is a commodity, and its value is determined by the amount of labor necessary for its production. If wages are sufficient to enable the laborers to sustain their families and reproduce their own labor power, then the exchange of labor power for wages is *not* unequal.

Where, then, does property income come from? To Marx, it originates in the workplace, in the transformation of labor power into labor. Quite simply, in the capitalist workplace, the workers do *more work than that necessary to reproduce their own labor power.* The difference between the labor performed by workers and that necessary to produce their labor power is *surplus labor* and produces ***surplus value.*** This surplus is the source of property income under capitalism.

To see where surplus value fits into the capitalist production process, it might help you to look at a circular flow diagram, such as Figure 1, which shows a continuous circulation of commodities measured in terms of their labor content or value—except for labor, which is the source of value. Start with the "Labor" box. At this point in the production process, direct labor works with material inputs to produce commodities. Some of these commodities go directly back into the production process to replace the material inputs used up in production. Other commodities go to the work force as wages. What is left over is surplus value.

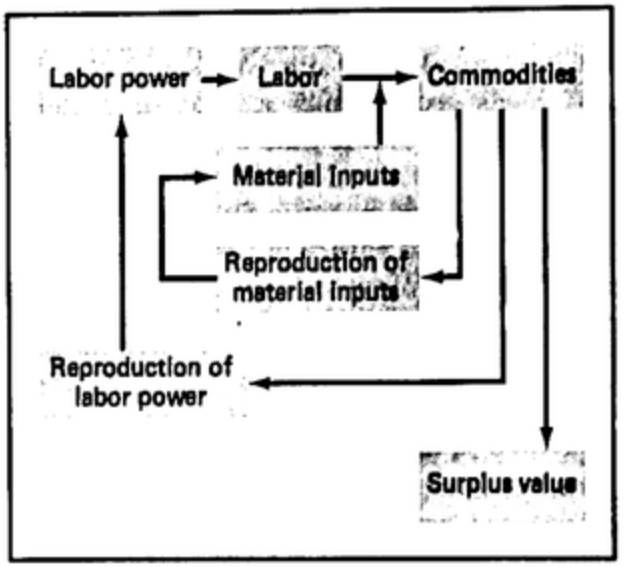


Figure 1 The circulation of commodities
In capitalist production, direct labor is combined with material inputs to produce commodities. Some of these commodities reproduce material inputs used up in the production process. Others, in the form of consumer goods, reproduce the capacity of the labor force to perform work—its labor power. What is left over is surplus value under the control of the capitalists.

Capitalism could not survive for long without surplus value. With no surplus, the entire economy would simply be reproducing its human and material means of production, with nothing left over for the owners of property. This might be possible in a commune or in an economy made up entirely of family farms and workshops, but it could not persist in an economy with large-scale private holdings of the means of production. With no surplus, there would be no consumption by the property-owning class. There would be no source of growth in the quantity of consumption goods and material inputs. Thus, a surplus is necessary to support the consumption of capitalists and to enable them to accumulate capital. Moreover, the commodity-producing industries must yield a surplus of material goods to feed, clothe, house, and transport service and government workers.

Of course, saying that capitalism produces a surplus to survive is like saying that birds grow wings so that they don't fall to the ground. It puts everything backwards. Capitalism doesn't produce a surplus because it has to. Surplus value is part of the very nature of capitalism. What needs to be explained is how and why surplus value arises in the capitalist mode of production.

Determinants of surplus value

If the product of one class of people is appropriated by another class, that is clearly ***exploitation.*** Under slavery, exploitation was direct and simple. The master controlled both the labor of the slave and the reproduction of his or her labor power. Limits on how much surplus labor could be extracted from each slave were imposed only by the slave's individual productivity and by the consumption necessary to keep him or her alive. Within these limits, the master had absolute control over the amount of surplus labor that the slave provided. To a Marxist, capitalist exploitation is much more subtle. The amount of the surplus is no longer determined by face-to-face encounters between worker and exploiter, but by properties of the economy as a whole. After all, labor power and commodities are exchanged on markets that are largely competitive.

To see what determines surplus value, consider how surplus value is calculated:

$$\text{Surplus value} = \frac{\text{(Hours worked per week)}}{\text{Labor}} - \frac{\text{(Amount of direct and indirect labor required to produce workers' consumption)}}{\text{Value of labor power}}$$

Under slavery, both labor and the laborer's consumption were determined directly by the slaveowner. Under capitalism, these same two items are determined in markets.

Such markets are rarely controlled by one capitalist, and the wages and hours emerging from the exchange of labor power depend on many factors. First and foremost are the historical traditions that limit the areas of negotiation. Current levels of wages and hours cannot depart too far from past levels without bringing wages and hours issues out of the marketplace and into the arena of open social conflict. As long as this does not occur, changes are largely determined by market conditions. When capitalism is profitable and unemployment is low, wages rise and the length of the work week falls. When capitalism is stagnant and unemployment is high, wages rise slowly, if at all, and the length of the work week drops very slowly. Prolonged prosperity tends to lower the surplus value that capital appropriates from each worker; prolonged bad times tend to raise it.

Another factor determining surplus value is labor productivity—how much labor is necessary to produce the commodities that make up the wage. Obviously, the more productive labor is, the smaller is the amount of work time necessary to replace labor power, and the greater is the surplus left for capitalists.

Productivity depends on two factors: how hard people work, and what kinds of tools they use. According to conventional economic theory, the most significant improvements in productivity are the result of technical progress. Yet, "technical progress" has taken three characteristic forms that make it seem anything but progressive to the people whose productivity is being raised. First is increased specialization or division of labor, to the point where mental and manual labor are divorced from each other. As a result, most jobs are repetitive and dull. Second is the increasing extent to which human labor is paced

by machines: Workers are forced to keep up with a machine that is now more like a taskmaster than a tool. Third is new machinery that displaces labor, so that progress may mean not easier work, but no work at all.

Some of the most bitter struggles between labor and capital have arisen over technical changes. In the late 1700s and throughout the 19th century, the introduction of machinery caused militant resistance by workers and violent retaliation from capitalists. *Luddism,* meaning sabotage by workers of the machinery that threatens their jobs, comes from this period. In the early 20th century, the vocabulary of technical change was enriched by the word *Taylorism,* the reduction of all manual labor to a series of preprogrammed, robotlike motions. This is often called the "deskilling" of labor. Efficient, perhaps, but hardly humanizing. In the late 20th century, we can expect increasing *automation* to wipe out entire occupations and to substitute mechanical robots for the displaced workers.

There is little doubt that the productivity increases that come from greater mechanization can have a lasting benefit for humanity. Marxists themselves have long thought of capitalism as a necessary stage in the development of human productive capacities. The *Communist Manifesto* of 1848, written by Marx and his collaborator Engels, contains some of the most effective capitalist propaganda ever written:

> The capitalist class, during its rule of scarce one hundred years, has created more massive and more colossal productive forces than have all preceding generations together. . . . What earlier century had even a presentiment that such productive forces slumbered in the lap of social labor?

But, Marxists believe, all this progress has been achieved at the cost of a perpetually insecure, alienated, and exploited class of workers. Remember, they say, capitalists have no reason to consider the interests of workers when they change the technique of production. They are driven by the search for profit and the need to survive competition from their fellow capitalists. Of course, even Marxists admit that rising productivity creates the *possibility* of rising living standards. But the *direct* consequence of technical change is usually a disaster for the working class. Historically, the trend in advanced capitalist countries has been toward a long-term rise in real wages, to the eventual benefit of the working class. But the process that brings it about is frequently destructive to the jobs, skills, and human capacities of this class.

The accumulation of capital

"Accumulate, accumulate! That is Moses and the prophets." In these words, Marx chanted the motto of the capitalist class. It expresses capitalism's innermost drive.

But ***accumulation of capital*** has two quite different faces. It consists, first of all, in the conversion of surplus value into material inputs of production, such as machinery, factories, and equipment, which are piled up on an ever-extending scale. This increase in productive capacity is not unique to capitalism. It is a characteristic of *any* growing economy. Second, however, a historical result of capital accumulation has been the spread of the capitalist system until it has come to dominate most of the world.

The class structure of capitalism

In its earliest days, "capitalism" was mainly confined to commerce. The forerunners of today's capitalists were traders, buying goods at one place or time and selling them at another. Most actual produc-

tion took place outside the primitive capitalist sphere.

Modern capitalism is quite different. It is primarily a *mode of production.* All modes of production combine and coordinate the ***forces of production***—labor power, material inputs, and technology—to produce and distribute their products. Various modes of production are so dramatically and obviously different because production takes place in a social setting. Production roles or tasks must be divided up. Who will decide what to produce and how to produce it? Who will provide the material inputs and who will provide the labor effort? Who will control the output? Even the most primitive societies assign some of these roles to different people. Each mode of production—slavery, feudalism, capitalism, or socialism—divides up these production tasks in a different way. The social setting of production is known as the ***relations of production.***

The ***class structure*** in capitalism is derived from its relations of production. One class of people, the ***working class,*** provides the labor power. Another class, the ***capitalists,*** provides the material inputs. Under the property laws and customs of capitalism, the capitalists who provide the material inputs decide what to do and how to do it. They also control the distribution of the product. The working class has the right to work or not, and the right to bargain over the terms for which it sells it labor power. But control over the production process and its product is exclusively the privilege of the capitalist class.

In the mid-1800s, it was fairly easy to analyze the class structure of capitalism, since the social relations of production were so simple. A typical capitalist enterprise was owned and managed by a single "master," who was its capitalist. Virtually everyone else connected with the enterprise was a manual worker, owning no capital.

Modern capitalist firms, however, are much more complicated than those of the 19th century. Technical advances have created a sizable corps of engineers and other production experts. The growth of national markets for certain commodities demands many office workers to engage in advertising, promotion, sales, and distribution. Large-scale industry has developed side by side with an entire managerial profession, trained in the major business skills. These new groups of technicians, marketing experts, and managers occupy an ambiguous position in the relations of production. They are dependent on wages or salaries like the working class, but like the capitalist class, they take almost no direct part in the production process itself. They are secure in their jobs as long as the firm itself is secure. Often, they identify themselves with the interests of their employers and think of their jobs as careers. Along with independent professionals and government officials, they make up what is usually known as the middle class. You probably come from this part of society and, together with most Americans, think of yourself as middle class.

But, Marxists are quick to point out, the two traditionally antagonistic classes of capitalism still exist. The owners of property and the top managers who control its accumulation make up the capitalist class, or ***bourgeoisie.*** And there is still a ***proletariat***—a class of manual workers who produce, transport, sell, service, and maintain the material basis for the rest of society. All other classes and groups depend on the working class for their needs.

Marxists believe that the main outlines of the capitalist class structure tend to be self-perpetuating: Since capitalists control output, they can replace its material inputs and money capital and can accumulate larger and larger quantities of both. The working class has a wage with which it maintains and reproduces its la-

bor power. But the surplus labor of the working class, which is the very source of capital accumulation, belongs to the capitalist class. Since the working class continually produces surplus value for capitalists but none for itself, it can never get control of the means of production. It remains dependent on the sale of its labor power for subsistence. This means, of course, that the working class is dependent on capitalism. It must participate in the reproduction of capitalism, or starve.

The tendency for capitalism to reproduce itself and to expand is not limited by national boundaries. The forces of capitalist accumulation have spread it to all corners of the world's economy, from its origins in Western Europe to every inhabited continent, even to the skies and waters of Antarctica. Today, it is difficult to find any industry that is not at least partly organized by capitalists. Yet, as forceful as this drive for capital accumulation is, there are limits to the rate of accumulation. One of the most important is the constraint placed on capital expansion by the rate of growth of the working class. The key link, Marxists argue, between the growth of capital and the growth of the working class is the pool of unemployed workers that is characteristic of capitalism.

The reserve army of the unemployed

Marxists call capitalism's unemployed population the ***reserve army of the unemployed.*** Some of the major influences on its size are shown in Figure 2. The cyclical expansions and setbacks of the capitalist sector have a big influence on the size of this reserve army. As capitalism expands in times of prosperity, workers are drawn out of the reserve army and put to work. When capitalism contracts during recessions, the reserve army is replenished with laid-off workers. Workers displaced by technical progress are also fed into it.

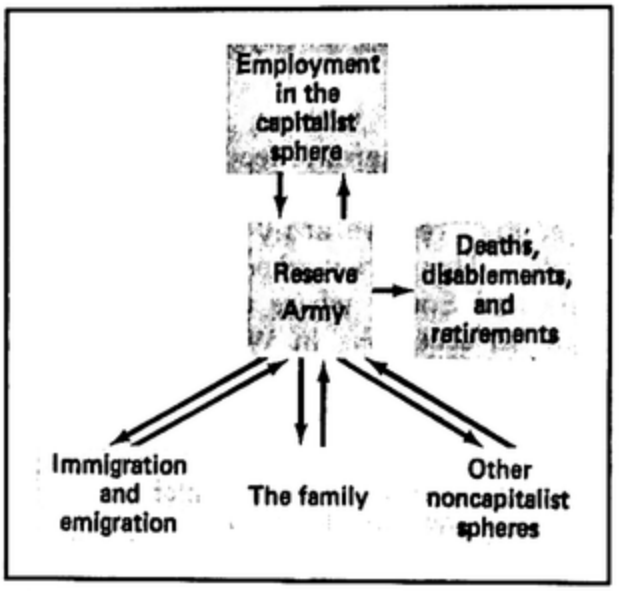


Figure 2 Flows of workers into and out of the reserve army of the unemployed

Changes in the size of the reserve army of the unemployed are determined by a complicated pattern of flows into and out of the various spheres of society. These flows are governed in part by the rate of capital accumulation.

The reserve army extends far beyond those normally thought of as the unemployed. Much of the reserve army are married women who take jobs outside the household when they can find them. When they lose their jobs, they go back to housework. Their unemployment never registers in the official statistics, which do not count people as unemployed unless they are "actively seeking work." But every woman who would like to work outside the home but cannot find a job belongs to the reserve army. Another source of entrants into the reserve army is the constant reduction in employment in handicraft trades, small retailing, and family agriculture. As capital accumulation destroys these modes of production, workers are released to join the reserve army. Yet another source of entrants is immigration, which provides recruits for the reserve army, while emigration reduces it.

The major types of flows into and out of the reserve army vary from one historical period to another. The rapid industrialization of the United States in the late 19th and early 20th century would not have

been possible without the enormous stream of immigrants feeding the reserve army. Today, West Germany has a large "standby" reserve army just south of its borders. These are workers who come from Italy and Spain, where unemployment is high, to seek jobs in West Germany. In good times, they are hired. In bad times, they are often sent home.

The reserve army, the accumulation of capital, and periodic crises

Marxists consider the reserve army of the unemployed important for two reasons. First, its size is a measure of capitalism's failure to provide useful work for the members of the very society it has created. The larger the reserve army, the greater the failure. Second, the reserve army provides flexibility for expansion. The larger the reserve army, the easier it is to accumulate capital without a labor shortage. Growth is easiest when capitalism is least successful in providing jobs for its workers. When capitalism does manage to put them to work, growth is difficult.

This sounds like a major ***contradiction of capitalism,*** and it is. But this contradiction has an important cyclical rhythm that sometimes brings the interests of workers and capitalists together, and sometimes brings them into conflict.

At the start of a business cycle upswing, both workers and capitalists benefit. Workers are more fully employed, and because they have larger incomes, they can buy more goods. The greater employment, output, and sales allow capitalists to reap more surplus value. But as expansion continues and the reserve army melts away, the interests of the two classes conflict. Workers are no longer "disciplined" by the threat of unemployment. They change jobs more often and work less intensively. They also demand higher wages. Lower productivity and higher wages mean lower profits. With less profit, the incentive and means for capital accumulation begin to dry up. The rate of capital accumulation drops off. Thus, Marxists argue, by successfully employing its labor power, capitalism has actually set the stage for a cyclical collapse or ***crisis.*** Now comes a period of contraction in employment, and the reserve army increases. Once again, the interests of workers and capitalists are parallel: The contraction is a disaster for both classes. But as the contraction continues, workers are again disciplined by widespread unemployment. Wage demands decrease and work effort increases, so that the value of labor power falls relative to work performed. Now profits begin to rise, and the prospect for profitable accumulation by capitalists starts to improve. Capitalism's very failure to provide jobs has set the stage for recovery.

You can see how important the reserve army is in producing this rhythm of crisis and recovery. As production expands, the reserve army declines, and capitalists and workers come into conflict. The conflict is resolved in favor of the workers, who do less work for higher wages. This slows down the rate of accumulation and produces a crisis. As production contracts and the reserve army grows, the conflict is resolved in favor of capital—workers produce more and earn less. The ultimate result of the contraction is an improvement in profit prospects, and a recovery can begin.

The contradiction that produces periodic crises is only one of the many contradictory tendencies of capitalism. Another is the growth of *unproductive labor* as capital is concentrated in large enterprises.

The concentration of capital and the growth of unproductive labor

The recurring rhythm of expansion and contraction that takes place under capitalism has an important effect on the pattern of accumulation. Because large enterprises are best able to survive in times of crisis,

the accumulation of capital has been coupled with the concentration of this capital into enormous and powerful firms. Massive corporations and capitalism go hand in hand.

The growth of these giant capitalist firms has had many important side effects. For example, think of the effect this concentration of capital has on the *type* of employees whom corporations must hire. Aside from the workers involved directly in the production of commodities, these firms need thousands of administrators to control the day-to-day operations of the corporations, and thousands more to operate their purchasing, record-keeping, and sales operations. This is quite a contrast to the capitalist enterprises of a century ago, in which a few hundred employees were easily managed by a capitalist and his clerk.

Another result of the emergence of giant corporations is the development of whole industries and occupations (such as air travel, telecommunications, corporate law, accountancy, and systems analysis) whose main purpose is to link together capitalist enterprises. A third by-product of the accumulation of capital on such a large scale has been the enormous growth of central government. To keep laws and regulations uniform and property secure, government grew right along with capitalist enterprises, both on a national and an international scale. Marxists think of the central bureaucracy, the judicial system, the diplomatic corps, and the military as a huge overhead expense of large-scale capitalism.

When you combine the numbers of capitalist employees and government bureaucrats who provide necessary support services for capitalist enterprises, yet who do no *direct* productive work, you begin to understand another important contradiction of capitalism. This "unproductive labor," however useful, must be supported out of the surplus labor of commodity producers. Thus, this ever-growing army of unproductive labor now competes with the very class that created it, the capitalists, for a share of surplus commodities. The larger the unproductive labor force, the smaller the amount of surplus available for capital accumulation. The fraction of the *working* population that provides the commodities that form the material basis for social life probably does not exceed 20 percent. All of the others must get their material means of existence from this 20 percent.

Of course, some unproductive labor would be necessary no matter how production and society are organized. It is hard to imagine a modern economy without doctors, fire fighters, clerical workers, public administrators, lawyers, teachers, and service workers. But capitalism—particularly the large-scale capitalism of the 20th century—has additional needs for unproductive workers: promotional and advertising workers, corporate lawyers, realtors, insurance agents, and countless others. Some people estimate that over 50 percent of GNP is devoted simply to the maintenance of the capitalist way of organizing the production process.

You can better understand the actual consequences of this unproductive labor by looking at Figure 3. The productive sphere of capitalism—the source of material commodities—gives rise to a surplus over and above what is needed to reproduce its material inputs and the labor power of its workers. Some of the surplus is accumulated as capital and remains within the productive sector. Some of it goes to the consumption of the capitalist class. But an enormous amount goes to reproduce the labor power of workers in the unproductive parts of capitalism, which include the many unproductive workers of commodity-producing firms themselves. Another share goes toward reproduction

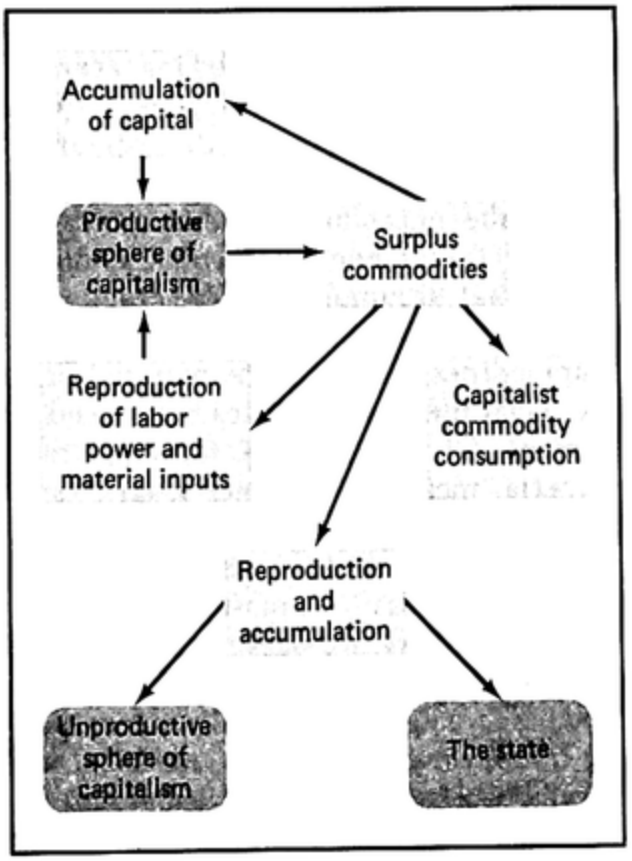


Figure 3 Productive and unproductive spheres of capitalism

Capitalism's unproductive spheres compete with its productive sphere for its surplus commodities.

and toward accumulating material inputs for the unproductive sphere. And a major share goes to support the government, or *state*, both to reproduce the labor power of its employees and to supply it with material inputs, such as aircraft, firehoses, pencils, offices, greenbacks, and red tape.

Thus, Marxists consider that a major contradiction of capitalism is its tendency to generate larger and larger amounts of unproductive work, which drains off an increasing share of output or surplus value. The very process of capital accumulation has set in motion forces that tend to slow down the rate of capital accumulation.

Paradoxically, though, the growth of the unproductive sector has in some ways helped capitalism to survive. After two centuries of development, capitalism has built up an enormous potential for producing material commodities. If the number of workers engaged in productive labor were larger, the result would be a great outpouring of material goods. Without a substantial change in income distribution, it is hard to see how this output could be sold at a profit. The most likely result of a larger productive sphere would be a glut of commodities and a crisis worse than the Great Depression of the 1930s. By draining off workers from the productive sector, the unproductive sphere helps to limit the production of commodities and to provide a market for those that are produced. Of course, it serves this function by performing a wide range of activities whose only social function is to preserve capitalism. To the extent that it does so Marxists argue, it must waste productive potential and limit capital accumulation.

Foreign trade and investment

A third contradiction that Marxists see as inherent in capitalism comes from its tendency to cross national borders. As major capitalist countries have searched for new markets for their output and new sources for their material inputs, they have developed trade links with one another and with so-called Third World (less developed) areas. The accumulation of capital has also meant a search for new ways in which to invest this capital. The result has been enormous flows of investment from one country to another.

As trade and foreign investment increase, national economies become much more dependent on one another. In particular, the capitalists of each country become committed to achieving a world order favorable to their own particular trade and investment patterns. Not surprisingly, rivalries among the capitalist nations quickly emerge, centering on control of world trade and investment. The result of these rivalries has been a series of wars

that have deeply wounded the capitalist world system. The existing world order broke apart following each of the two World Wars, and produced giant socialist antagonists, the Soviet Union and China. Various other parts of the capitalist system broke away and established themselves as independent, often socialist, countries.

Even when the international expansion of capitalism was at its peak, nearly a century ago, it inflicted enormous burdens on the capital accumulation process. Empire building benefited some sectors of the capitalist system, but was very costly to others. The maintenance of armed forces to defend colonies, trading partners, overseas investments, and trade routes was a continual drain on the flow of surplus commodities from the productive sector. Thus, Marxists believe, the capital accumulation process was in yet another way self-destructive.

Contradictions and the collapse of capitalism

You have now learned about three of what Marxists consider to be the major contradictions of capitalism:

1. Expansion of capital requires expansion in employment. When the reserve army of the unemployed is depleted, the accumulation process is brought to a halt, and a crisis follows. Thus, expansion sets to work forces that produce a contraction.
2. Accumulation of capital leads to its concentration in enormous firms operating on a national or world scale. Although such firms are immensely profitable, their operation requires large numbers of unproductive workers internally, in supporting industries, and in the state. These unproductive activities divert large quantities of surplus commodities away from the accumulation of capital.
3. Capital accumulation on a world scale produces colonialism, militarism, and war. Warfare among the capitalist countries is destructive of capital and it limits further accumulation. Wars have also led to the breakup of the capitalist world system and the establishment of socialism in a large part of the world, thus limiting the world sphere of capital.

You will notice that in each of these cases, Marxists argue that the very process of capital accumulation cripples itself. Accumulation leads to nonaccumulation. This unity of opposites is characteristic of all historical processes. Marxists consider it the mechanism of all fundamental social change. They call it the ***dialectic of history.*** Every contradictory mode of production contains within it the forces of its own destruction.

For a long time, Marxists thought that capitalism would almost automatically self-destruct from its own contradictions. Marx himself put it this way in *Das Kapital:*

> Along with the constantly diminishing number of the magnates of capital. . .grows the mass of misery, oppression, slavery, degradation, exploitation, but with this too grows the revolt of the working-class, a class always increasing in numbers and disciplined, united, organized by the very mechanism of the process of capitalist production itself. The monopoly of capital becomes a fetter upon the mode of production, which has sprung up and flourished along with, and under it. Centralization of the means of production and socialization of labour at last reach a point where they become incompatible with their capitalist integument. This integument is burst asunder. The knell of capitalist private property sounds. The expropriators are expropriated.

This prediction was naïve and inconsistent with some of Marx's other beliefs about the ability of humanity to control its own history. In particular, it neglected the importance of politics and its relationship both to the preservation and to the transformation of the mode of production.

Summary

Some of the more important Marxian ideas are:

1. Productive work is the characteristic activity of humanity. Under capitalism, work is organized around the production of commodities, which are material objects produced for market by wage labor. These commodities embody all the labor that has gone into them. Under capitalism, the use to which commodities are put determines indirectly the use to which labor is put.
2. The value of a commodity is the amount of labor that goes into its production. This value may be broken down into two parts. The first is the value of the commodities that are necessary for the reproduction of the laborer. This is known as the value of labor power. Since this is smaller than the labor done by the worker, there remains a surplus over and above the value of labor power. This is called surplus value. It is the source of all property income arising from capitalist production.
3. Surplus value comes from exploitation of workers. Labor produces everything. But since workers put in longer work days than would be necessary to produce their consumption goods, they perform surplus labor. Their surplus product becomes the property of the capitalist employer.
4. This surplus product mainly goes to the accumulation of capital, the material means of production and wage goods for the maintenance of workers. This is the source of capitalism's expansionary power.
5. Capital and labor power are the forces of production. Since capital is owned by one group of people and labor power by another, the forces of production are mirrored in the social relations of production. One group owns and controls. This is the capitalist class. Another neither owns nor controls, but has only the capacity to work. This is the working class.
6. The working class is divided into those who are employed and those who are not. The reserve army of the unemployed is one of the main mechanisms for the preservation of capitalist social relations, since unemployment is the main weapon that keeps the working class under control.
7. Capitalism is subject to contradictions, or self-destructive tendencies. One of these involves the accumulation of capital, expansion of production and employment, and depletion of the reserve army. Capital temporarily loses some of its power over labor. Productivity falls and wages rise. The result is a drop in surplus value and accumulation, known as a crisis. The crisis slows the growth in employment, replenishes the reserve army, and restores the conditions for capitalist control.
8. Capital accumulation has other contradictory tendencies. Among them is the growth of large enterprises whose administration and sales activities absorb large numbers of workers who produce no product. Another is a tendency to spread across national boundaries and to require substantial diplomatic and military support. The maintenance of such unproductive activities absorbs a large portion of the surplus product and impairs capital accumulation.

Key concepts

Conventional wisdom

Commodity

Capitalist mode of production
Value
Capital
Labor and labor power
Surplus value
Exploitation
Accumulation of capital
Forces of production
Relations of production
Class structure
Working class
Capitalists
Bourgeoisie
Proletariat
Reserve army of the unemployed
Contraditions of capitalism
Crisis
Dialectic of history

Questions for review

1. Why is the distinction between labor power and labor crucial to understanding where surplus value comes from? Why do workers willingly produce a surplus over and above the value of their labor power?
2. "The accumulation of capital is the spread of capitalism across the face of the earth." Discuss.
3. Do you consider yourself a member of the capitalist class, the working class, or the middle class? Why? If you have a job, do you own the means of production you work with, or does someone else own them? Do you control what happens to your product?
4. What is a "contradiction"? Can you think of any contradictions of capitalism that are not discussed in this chapter?
5. What fraction of the people you know are actually productive laborers? What about your economics teacher?
6. Do you expect the capitalist system to break down from its internal contradictions?

38

The Economies of Russia and China

As you read and study this chapter, you will learn:

- how the Russian economy developed into one of the major industrial economies of the world
- how the Russian planned economy functions
- why the Chinese pursued a development strategy different from that of the Russians
- why Chinese policy has sometimes emphasized economic development, and sometimes social development

Most of us react positively to the idea of natural foods, even though we may not much like soybeans. There is a long tradition in our culture that treats things that are "natural" as though they were inherently good. It goes back at least to the biblical account of the Garden of Eden.

Of course, most foods that come from nature taste terrible. Wild animal meat can be stringy and harsh. Wild fruits and vegetables are often dry and insect ridden. Nearly all the foods that we think of as "natural" have benefited from many centuries of human cultivation. We have improved the genetic stock of both plants and animals and learned how to make the most of their genetic potential. Only fish seem to do better on their own. The next time you enjoy your crunchy granola, remember that nearly everything good in it is a human creation, no matter how free it may be of the products of the chemical industry.

The great classical economists such as Adam Smith firmly believed that there are laws of nature regulating social life as

well as the physical universe, and that societies would function well if their institutions were harmonious with these laws. Since they were also champions of developing industrial capitalism, it is not surprising that the classical economists thought of capitalist institutions as natural and good. To this day, our thinking and language about market economies is colored by their attitude. Price controls, rationing, import quotas, and acreage allotments are often spoken of as "artificial" restrictions on the functioning of markets. Since the opposite of *artificial* is *natural*, this kind of language treats the free market as a natural and, by implication therefore, a good institution.

The institutions of modern capitalism, however, are actually no more natural than crunchy granola. Adam Smith thought that people had an innate tendency to swap or exchange goods, "a propensity to truck and barter." From this, he concluded that the division of labor and production for the market were extensions of a natural human trait, and were themselves natural. But what is so natural about General Motors and the New York Stock Exchange? What do they have to do with people's propensity to truck and barter? These giants did not just spring up like mushrooms in wet weather. Capitalist institutions are the result of many decades of purposeful human effort to improve and refine the division of labor and production for profit. They have only been dominant in the past two centuries. After studying the alternatives, you may conclude that capitalism is superior. But it is not more natural.

To study precapitalist modes of production, you will have to go to the history and anthropology departments. But the study of contemporary alternatives to capitalism is very much a part of the economics curriculum. Comparative economic systems is one of the main branches of applied economic analysis. Your own economics department probably offers survey courses in this field and may also have specific courses on the economies of the Soviet Union and the People's Republic of China, to which this chapter is devoted.

The Soviet Union arose during World War I from the wreckage of Tsarist Russia. After a decade of experimenting with various development strategies, the Soviet leaders began to pursue a program of centralizing economic power in the hands of the government and concentrating on heavy industrial development. Since the Russian development strategy was the first adopted by a Marxist country, it is the most obvious place to start the study of socialist economies.

Economic planning: The Soviet economy

The Soviet Union has a ***centrally planned economy.*** Of course, all economies involve some planning. To take an example from Western economies, protective tariffs and subsidies are conscious efforts to guide resource allocation and industrial development. But their impact is indirect. Private enterprises still make production decisions, but in a context that includes tariffs and subsidies. In a centrally planned economy, a planning bureau makes the major production decisions directly. The enterprises are expected merely to carry out the plans of the bureau to the best of their ability. Their managers are executors, not decision makers.

The institutions of central planning in the Soviet Union, like the institutions of capitalism, have been molded by the process of development. To understand them, it helps to know some history.

Some historical background

On the eve of World War I, Tsarist Russia was a study in economic contrasts. It had

a small but quite modern capitalist industrial sector, yet 80 percent of the population tilled the land. Russian peasant agriculture, only half a century after the emancipation of the serfs, was very backward, although it did produce an abundance of grain. The peasants were, for the most part, desperately poor. By contrast, the cities were very prosperous, and a rich scientific and cultural life flourished in them.

In 1914, at the very start of the war, the Russian army was badly beaten by the Germans. Nevertheless, Russia fought on for three more years. This continuation of the war merely served to weaken the economy and the Tsarist government further. It set the stage for the Revolution of 1917. In that year, the Communist party (then known as the Bolshevik party) seized state power from the left-wing republicans who had replaced the Tsar and gained control of industrial centers and some of the countryside. It made peace with Germany, but immediately was faced with a civil war against armies loyal to the old order. There was also an armed intervention by troops from several countries, including the United States, so that it was not until 1921 that the Bolsheviks gained full political control. Seven years of external and internal war had taken a fearful toll.

During the war period, the Soviet government took over most industry, banking, and foreign trade, and tried to run the economy through a combination of directives, coercion, and patriotic appeal, known as *War Communism*. This approach grew less and less successful as time passed, particularly because of difficulties with the food supply. Because of the war, the government had few manufactured goods to offer the peasants in exchange for food. The peasants, understandably, refused to supply food for paper money that could not buy what they wanted. They just stopped producing. The system of War Communism finally broke down in 1921. It was replaced by Lenin's *New Economic Policy*. This was a reversion to the market economy in most areas (other than industry, banking, and foreign trade). Individual profit incentives worked in their customary way, and both industry and agriculture recovered from the war and the effects of War Communism. By 1928, the output had reached levels roughly comparable to those of 1913, before the outbreak of the war.

Stalin's development strategy

At the time of the 1917 Revolution, the Bolshevik leaders had expected successful socialist revolutions to break out throughout the world. When this didn't happen, Russia found itself isolated, backward, poor, and militarily weak. The political and economic situations both called for a program of rapid development. During the mid-1920s, there was a long debate among top party officials about the right strategy to follow: whether to favor industry, to concentrate on agriculture, or to strive for balance between the two sectors. The debate was settled by Stalin, who gained uncontested political power in 1928 and launched a program of ***forced industrialization.***

"Forced" industrialization means a rapid rate of growth in the industrial sector relative to other sectors of the economy. Since every industrial economy was once an agricultural economy, every one went through a period of forced industrialization. Socialist and capitalist countries differ only with respect to how the forcing is done. In Russia, as we will see, it was particularly brutal.

All rapid industrialization is subject to a set of common constraints imposed by the relationships of inputs and outputs:

1. The industrial labor force must grow rapidly.

2. It must be fed, and the food cannot come from industry itself.
3. The growing industrial labor force must be equipped with material means of production, which must be provided from somewhere.

Russia's isolation from the rest of the world implied some further constraints:

4. Since there was no immigration, the growth in the industrial labor force required internal migration from farm to city.
5. Since there was no importation of food, the industrial labor force had to be fed by domestic agriculture.
6. Since domestic industry had to provide its own material means of production, agriculture could not receive much industrial output in exchange for its foods.

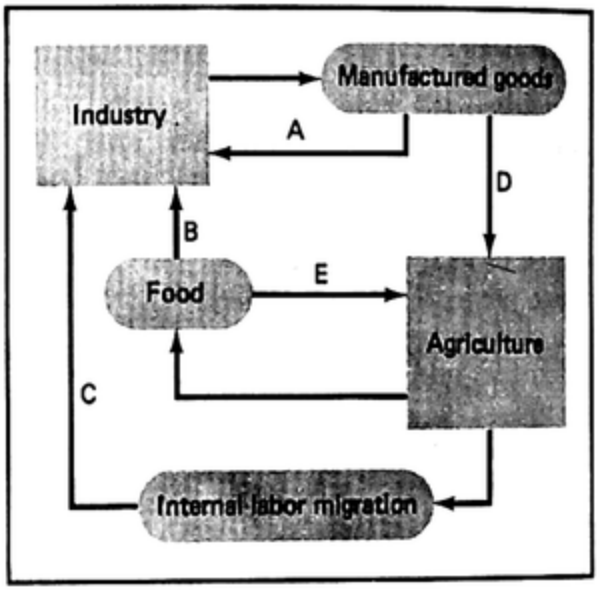


Figure 1 The relationship between industry and agriculture

Rapid industrialization requires a large flow of labor and food from agriculture to industry, and retention of most industrial output in the industrial sector itself. Thus, Flows *A*, *B*, and *C* must be large, and Flows *D* and *E* must be small. This implies that agriculture must be squeezed to provide net resources for industry.

These relationships are spelled out in Figure 1. If the industrial sphere is to grow rapidly, it must have large inputs of manufactured goods, food, and new workers (Flows *A*, *B*, and *C*). This means that agriculture must supply food and workers. It cannot retain its food or absorb a large proportion of manufacturing output, so Flows *D* and *E* must be small. Forced industrialization must come at the expense of the agricultural sector.

The Stalinist solution to this problem of unequal flows between industry and agriculture was the ***collectivization of agriculture.*** This was the most brutal episode in the entire Stalinist era, which was notorious for its brutality. The peasants were forced by violent means to consolidate their landholdings into *collective farms*, to farm them in common, and to sell most of their output to the government at very low prices. They could not keep much of what they produced or get much in return. Predictably, they resisted. They refused to plant crops, slaughtered their animals, and hoarded what surplus food they did produce. Millions starved to death or were killed. The productivity gains that were expected from merging small holdings into large collectives didn't materialize because the peasants had so little incentive to work hard. Productivity was disastrously low in the early years of collectivization. This, of course, kept the industrialization program from working as well as it otherwise might have. But Stalin persisted. After agriculture recovered from the initial shock of collectivization, the system functioned about as planned. The countryside was mercilessly squeezed to provide food and workers for the industrial labor force. Despite an enormous cost in human misery, as an institution for reallocating resources, the collective was a success. But as a production institution, it has never worked well because it gives so little incentive to the peasants themselves.

Collectivization provided the means for industrial development, but these means had to be coordinated to achieve the desired effect. This was done by setting up a planning bureau charged with drawing up a blueprint for growth, and establishing a large bureaucracy to implement the resulting plans. Since there was no viable historical precedent for either drawing up coordinated plans or running an entire economic system by directives, the initial planning efforts were chaotic. Results did not correspond to the plans, partly because the plans were not very realistic, and partly because the bureaucracy was simply not capable of carrying them out.

The major planning instruments were a series of *Five-Year Plans*, the first of which was issued in 1928. These were expressions of the Communist party's overall development goals, as the economic planners expressed them in quantitative terms. They set target growth rates for a wide variety of products whose production processes were tied together by the input-output structure. You can appreciate how complex this kind of planning must have been for people who were new to the game. This was particularly true because the overall plans were very ambitious.

Year-to-year production decisions were guided by short-term plans that governed the operations of the enterprises. Since these had to be specified in great detail, preparing them required far more expertise than the inexperienced planners could hope to have at their disposal, and many mistakes were made. Nonetheless, the growth rate of Soviet real GNP during the first two Five-Year Plans seems to have been somewhere between 5 and 12 percent per year, depending on how various products are weighted. This was remarkable, especially when you consider the disastrous human and economic damage brought about by collectivization.

World War II, much of which was fought on Russian soil, dealt the Soviet Union another painful blow. Taking into account both the war-caused higher death rate and lower birth rate, 40 or 50 million Russian people were lost in this war. Production did not regain its 1940 level until about 1948. Over the same time span, the U.S. real GNP increased about 50 percent.

After the war, the Stalinist strategy was resumed, and growth rates remained high. Following Stalin's death in 1953, however, the program of forced industrialization was gradually relaxed in favor of higher output of consumer goods and higher living standards, especially in the agricultural sector. The Soviet Union today is a very different place from what it was during the early period of the Five-Year-Plans: far more modern, with a higher living standard. And it is less brutal than it was during the collectivization period, although political dissidents continue to be treated very harshly. But the major economic institutions of Soviet society are still much like those that were developed to achieve forced industrialization.

Soviet economic structure

You have learned to analyze the structure of the American economy by breaking it down into four sectors—households, firms, governments, and foreign trade—and studying the linkages among them. These linkages mainly take the form of market relationships and voluntary exchange (except, of course, for the collection of taxes). The participants are guided by their own self-interest and by the economic laws of the system within which they operate. But in a planned economy, most economic activity is governed by directives rather than by market incentives. The economic analysis that helps you understand the market economy cannot be applied in a simple way to the study of a planned economy. To

understand how such an economy functions, you must study the constraints and incentives acting on the sectors as they operate in a context of planning.

Households The household sector in the Soviet economy is similar in several ways to its American counterpart. It receives much of its income in the form of money, and buys many of its consumption goods with this income. As wage earners, household members can change jobs for higher wages or more agreeable work. As consumers, they can exercise their preferences among the very limited range of goods available to them. Peasants living on collective farms have small individual plots of ground on which they can grow food when they are not working on the collective land. They can either consume their products or sell them on a market. In these respects, household members live in a market economy and are regulated in their economic behavior by the incentives of a price system.

But two important institutional facts distinguish the *economic* lives of Russian citizens from those of citizens in the West:

1. To a far greater extent than is true in most Western countries, Russians get certain kinds of consumption goods without having to pay for them. These range from medical care, education, and old-age subsistence, which are available to all, up to limousines and fine houses, which are given only to officials and managers.
2. Since the means of production are owned by the state on behalf of the whole nation, individuals cannot accumulate ownership of capital. They may save in state banks and buy state bonds, but they cannot found private financial enterprises.

Because of these facts, there is little incentive for private saving. Most saving in the Soviet Union is carried out by the state, in the form of tax revenues and profit margins of state enterprises.

Firms Russian enterprises are very different from their Western counterparts, first because they are publicly owned, and second because of how planning directives limit the behavior of their managers:

1. The state sets wages and prices for nearly all commodities. Market conditions influence wages and prices only through their effects on planners, not through their effects on firm behavior.
2. Planning directives establish output goals for firms, and material allocations limit their input choices. Thus, the managers operate within a very narrow range of choices.
3. Output goals are largely set in terms of quantities—so many kilos or meters of output. Managerial bonuses and promotions are mainly tied to meeting these quantitative goals.

Most Western economists are very critical of the constraints imposed on Soviet firms. They argue that Russian resource allocation would be far more efficient if the managers of firms could adjust prices in response to supply-and-demand conditions, if they could freely choose their outputs and inputs, and if their goals and incentives were tied to profitability rather than quantity. They think that then there would be far less waste of labor and material inputs, and far more attention to the quality as well as the quantity of output. The "reformers" within the Soviet economic profession make similar arguments and have had some effect on practice. But since such fundamental changes in the

structure of the Soviet firm would greatly redistribute and decentralize decision-making power, there is strong bureaucratic resistance to change.

Foreign trade Trade with other countries is a state monopoly in the Soviet Union. The content and direction of trade is far more subject to the climate of international affairs than the trade of Western countries is. This results, of course, from the fact that party policy directly controls how the state monopoly buys and sells. Yet, the direct needs of the Russian economy also influence its international trade. The Soviet Union is willing to trade its strategic mineral products to the West in exchange for high-technology capital goods. And when the Russian grain crop fails, as it often does, the Soviet Union becomes heavily dependent on the Western agricultural market to feed its population.

Government The government sector of the Russian economy is far more controlling than Western governments. In the United States, the government sector taxes, buys, and regulates. In health, education, and public utilities, it competes with private industry. It has a monopoly over national defense and domestic police powers. But it has only limited control over prices, wages, resource allocation, income distribution (before taxes), and capital accumulation. The Soviet government does all that and much more besides. It monopolizes banking and foreign trade and owns the material means of production. Thus, it exercises all the functions that in our country are performed by the private and the government sectors combined, and it does so with very limited attention to consumer wants. When you started to study the U.S. economy, you learned how a two-sector economy—households and firms—might work. It would be silly to approach the Soviet economy in the same way. To understand how the Russian economy functions, you must start with the government sector and its planning process.

The planning process

In the West, the fundamental economic problems of what, how, and for whom to produce are solved in a decentralized way. Millions of people and thousands of firms pursue their own individual goals. No one plans what happens to the economy as a whole. Of course, the government intervenes to influence aggregate outcomes, like the rates of unemployment and inflation, but the fundamental economic problems of what, how, and for whom are not solved by anyone. They are solved by a social system.

In a centrally planned, socialist economy, the decisions as to what, how, and for whom to produce are made in a far more deliberate way. The government or ruling party sets social or political goals reflecting its view of the best interests of the country. To a far greater degree than is true in the United States, what happens in a socialist economy is the result of conscious choice.

In Russia, the goals are established by the *Politburo* of the Communist party. Its leaders represent themselves as the personification of the will of the Russian working class, as the *vanguard of the proletariat.* Yet, it is often argued that the party leadership simply makes up a new ruling class, by virtue of its control over the means of production. Since Stalin's time, the party's economic goals have consistently emphasized rapid industrialization and the maintenance of a large and modern defense system. The provision of a higher living standard has usually been

much less important, although consumer goods have been given more attention in recent decades, largely because of the success of the program of forced industrialization.

The agency responsible for drawing up overall national plans is the State Planning Committee, or *Gosplan*, as it is usually called. At its top levels, Gosplan is staffed with people who work with high party and government officials. At lower levels, it is staffed with economic and technical experts.

Gosplan has three different time horizons. It draws up 10- to 20-year projections for population, labor supply, technological change, transportation, natural resources, and other major variables. The five-year plans express the broad outlines of economic policy for the periods they cover. One-year plans spell out the concrete steps by which the goals of the five-year plans are to be implemented. Only these annual plans have direct significance for the enterprises.

Between the detailed annual plan and the specific goals established for each of the several hundred thousand industrial, agricultural, and distributional enterprise units, there are many layers of decision making and technical implementation. An enormous bureaucracy with crisscrossing lines of authority over industries and regions is necessary to turn the plan into fact. The bureaucracy must coordinate not only physical quantities of inputs and outputs but also wages, prices, and financing. The most telling criticism of the Soviet economic system is that it cannot function efficiently because the problem of coordination is just too complex to be solved by even the best bureaucracy (and the Soviet bureaucracy is, in many respects, far from the best).

To see why, think for a bit about what is involved in working out a detailed plan. First, suppose that production were much simpler than it is. If consumption, investment, and government goods were produced by unskilled labor alone, the planners could set targets for various outputs, calculate the labor required to produce them, estimate labor supplies in various parts of the country, and reach some conclusion about the *feasibility* of their targets. If the targets seemed too ambitious or too modest for the labor available, they could be modified to meet the realities of *production possibilities.*

One level of complexity is introduced by the fact that resources have very specific capabilities in the short run. People who know how to mine coal cannot learn how to operate a steel furnace overnight. Nor can steel be melted in a coal mine. The options open to planners are thus really limited in the short run. It is not just *any* labor, but the properly skilled labor that must be available in the right quantity in the right place.

Another complexity comes from the interdependence of the sectors of the economy. If a plan for coal production is overly ambitious, the actual output of coal will turn out to be far less than anticipated. Thus, less coal will be available for steel production; less steel will be available for machinery production; and less machinery will be available to produce other goods. A shortfall in one sector seriously impairs the plans or targets of other sectors.

The overall plan must be *consistent* with available resources and with the interdependencies of production. Not surprisingly, much of the effort of the Soviet planners is aimed at achieving consistency. *Materials balances*, which are input-output accounts for about 20,000 products, are carefully calculated and plans are modified to achieve consistency. Russian economists and mathematicians have worked hard in recent decades to learn how they might efficiently balance inputs and outputs with the help of computers,

but the sheer size of the problem has so far kept them from finding a truly efficient solution. A 20,000-sector input-output table requires 400 million entries. Failure to achieve consistency means shortages, surpluses, excess capacity, and unwanted inventories, so that the Russian economy is chronically *inside* its production possibility curve rather than on the *frontier*. Even Russian economists admit this.

However, the Soviet system has grown at a rapid rate. Even though it wastes much of its potential, that potential shifts outward rapidly, dragging actual output along with it. This is illustrated in Figure 2. In the long run, the Russians may be better off with what they have than they would be if their economy were more efficient, but grew more slowly.

Despite the enormous complexity of planning, the Soviet economy does work. One of the main reasons it does is that change is gradual. Even a rapidly developing country changes little from one year to the next. The planners start from a knowledge of what has worked in the past, and grope toward where they want to be in the future. The continuity of behavior and technology make it possible to do fairly well, even though mistakes are frequent and costly.

Prices

In Western economies, prices provide market signals that guide resource allocation, at least in those sectors that are fairly competitive. To a limited extent, this is also true in the Soviet Union. About 10 percent of agricultural output is exchanged on the collective farm market. It is produced by collective farm workers on small individual plots that they may use to grow crops for their own use or for market sales. This is not really capitalism, since the workers do not own the land and they cannot employ others to work for them. But prices on the collective farm market fluctuate freely in response to changing supply and demand, and the farmers are guided by these prices in their choice of what to produce.

Most prices are set by the state, however. They are not arbitrary, but are estab-

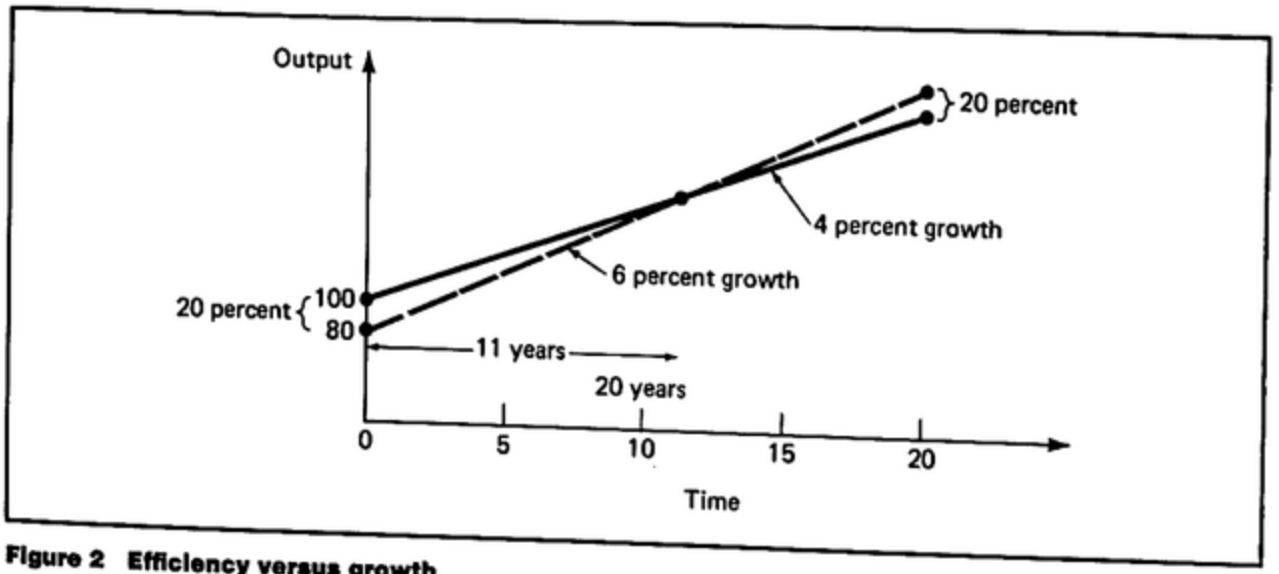


Figure 2 Efficiency versus growth

In 11 years, an economic system that is only 80 percent efficient but grows at 6 percent a year will catch up with one that is 100 percent efficient but grows at only 4 percent a year. In 20 years, it will have a 20 percent higher output than the efficient, slow-growing system.

lished with specific goals in mind. State planners tend to set low prices on necessities, such as meat and housing, regardless of the amount available relative to demand. The result is what you would expect: Although everyone can afford such goods, not everyone can get them. There are chronic shortages, since the quantity demanded at low prices greatly exceeds the supply. For many goods, however, the price set by the planners is meant to be that at which consumer demand will equal planned output. Yet, if either supply or demand differs from the planners' expectations, shortages or surpluses will develop. Since the market price cannot deviate from that set by the state, the shortages and surpluses will persist rather than be eliminated by price adjustments, as they would be in a competitive market.

The gap between production costs and market prices is filled mainly by a kind of sales tax called the *turnover tax*, which does two things. First, since it is levied on nearly all goods, it reduces purchasing power. The turnover tax performs the same function as the income tax does in the United States: It siphons off purchasing power, so that consumer income will not command resources of greater value than those devoted to the production of consumer goods. Second, since it is levied at higher rates on some goods than on others, it discourages the consumption of goods that the planners wish to produce in small quantities because of the particular resources they draw on. Thus, the rate structure of the turnover tax is designed to coordinate consumer demand and the production plan. The failure to achieve this coordination well enough to prevent shortages has been responsible for at least part of the Soviet consumers' dissatisfaction with their living standards, although their main complaint seems to be about the shoddiness of the available goods. This shoddiness results largely because production targets are specified in terms of quantity rather than consumer acceptability. Producers can get away with low quality because so many goods are in chronic short supply.

Another area in which supply and demand must be taken into account is in the setting of wages. Since Russians are more or less free to change jobs, rapidly expanding industries must be able to attract labor if they are to meet their production targets. A failure to adjust the interindustry wage structure in response to changes in output targets would produce shortages and surpluses of labor. The planners must use relative wages to influence the pattern of labor supply to various industries.

Note that in all these cases in which prices must be set with supply and demand in mind, at least one of the parties is a household, which has a range of choice in the Soviet Union that enterprises do not have. Enterprises receive both output targets and input allocations. Since they cannot choose their inputs according to price, the prices of industrial products traded between state enterprises do not have to equate supply and demand. Resource allocation is determined by planning alone. Yet, input prices do have an important influence on how the economy operates. Industrial products are priced at industry-wide average costs of production, including an allowance for the use of fixed capital. Firms that use their input allocations more productively than average show a profit. Since bonuses and other incentive payments are partly related to profitability, this pricing practice encourages more efficient use of inputs.

Income distribution

The structure of incomes in the Soviet Union reflects the widespread use of ***material incentives*** to encourage productivity. Different industries have different wage

scales to assure labor mobility toward those sectors whose growth is planned to be higher than average. Within enterprises, there is an elaborate system of wage differentials and bonuses to reward good performance. And high officials, artists, and athletes are paid on a scale that compares with that in the West. The Soviet Union is far less egalitarian than many people might think. In fact, its distribution of wage and salary income is probably not much different from that in the United States. But Russia has no capitalists, whose large property incomes create the extreme upper end of the income distribution in Western countries.

Is Russia a socialist country?

Russia is clearly not an egalitarian society, but is it even a socialist society? You may argue that it must be, since it has very little that looks like private enterprise. But a left-wing socialist, influenced by the thinking of Mao Tse-tung, would argue that the extent of bureaucratic hierarchy and the use of material incentives make Russia capitalist!

Marxist and left-wing thinkers of various shades of opinion often define socialism to match their own view of how society ought to be organized. Most arguments over whether Russia is actually socialist turn out to be arguments about the definition of socialism.

In his book *Post-Revolutionary Society*, the American Marxist Paul Sweezy argues that while the original Bolsheviks were the "vanguard of the proletariat," both they and their proletarian supporters were greatly reduced in numbers by the 1917–1921 civil war. To try to govern the country despite their own weakness, they encouraged the development of a corps of state officials. These officials did not share the Bolsheviks' revolutionary goals, however. Stalinism, which further developed the planning apparatus and purged the remaining old Bolsheviks, only completed a process that was already under way. With the end of Stalin's personal rule, these state officials came to make up a ***new ruling class.*** They control the means of production by controlling the state and party, which control the planning apparatus. They are not capitalists, since they do not personally *own* the means of production, but they function like a capitalist class.

Despite the power and privilege of the Russian ruling class, and the fact that entry into it is very difficult for those who are not the children of ruling-class parents, it seems to be viewed apathetically by the mass of workers and peasants. In fact, Russian dissidents largely come from the ruling class itself. Workers and peasants are either satisfied with their guaranteed employment, high social wage, and (in recent decades) rising living standard, or are unable to express their dissatisfaction. Perhaps because they feel they lack the means to change the system they live under, they are not revolutionary.

Indeed, because the ruling class has such tight control over the political process, they are not even very political. Workers have little interest in the goals of the ruling class, since they have no role in formulating national goals. In the long run, this is a source of national political weakness. A society in which the great mass of the people are apathetic and self-absorbed is in trouble, whatever its economic system.

The People's Republic of China

China is a unique and colorful country. But its extraordinary population problem would command your attention even if there were nothing else interesting about its economic system. China's population is

roughly a billion people. Every 15 years or so, it adds to its population about as many people as there are in the whole United States. Although the Chinese land area is large (about the same as that of the United States) much of it is unfit for farming, and the ratio of population to arable land is among the highest anywhere on the globe. Intense cultivation makes this land very productive per acre, but it is not very productive per worker.

In some respects, the Chinese economy is similar to the Russian, but in others, it is quite different. You already have some understanding of how the Soviet economy developed and how it now differs from a Western economy. This part of the chapter will compare China and the Soviet Union to give you a better picture of the variety of forms that socialism can take. First, however, it will help to review some Chinese political history, since the particular form that China's socialist development has taken has been shaped by that history.

The Chinese Revolution

Chinese civilization has existed for scores of centuries, but China has traditionally been hard to govern because of its mountainous terrain, which divides it into numerous separate areas. During the late 19th century, the coastal areas were easy targets for European and Japanese imperialists who wanted to set up trading concessions. After the downfall of the traditional Chinese empire in the Revolution of 1911, led by the Kuomintang (or Nationalist party), the interior was divided up into areas ruled by various warlords.

Gradually the country began to be unified by the Kuomintang. The Communist party, which was formed in 1921, cooperated with the Kuomintang in its early years. In 1927, however, this alliance was broken, and the Communist party retreated into the rural interior and began its own revolution. For much of the 1930s and 1940s, the Communists fought both the Kuomintang and the Japanese, who began their invasion of north China in 1937. But the mountains of China's interior provided the Communists with protection, and their revolutionary movement gathered strength. It was led by Mao Tse-tung, who continued to lead the Communist party after it drove the remnants of the Kuomintang to Taiwan in 1949 and took over the governance of the country.

An extremely important feature of the Chinese Communist party was its basis in the peasantry. While the Russian Bolsheviks drew their numerical strength from the urban working class, Mao's forces didn't occupy the cities until after the revolution. They spent the 20 years of revolution in various agricultural areas in the interior. There they took over the administration of the government, fought the landlords, subdivided the land, and developed primitive socialist societies. By the time they took over the entire country, the Chinese Communists had had much experience in administration. Party membership had also grown enormously during the years in the interior. By 1945, when the Japanese were defeated and full-scale civil war against the Kuomintang began, the party had over a million members. By 1949, the victorious party had roughly 5 million members, about 80 percent of them peasants. This peasant base gave the Chinese Revolution its distinctive character. Agriculture and the peasantry have always been of central importance in post-revolutionary economic development both because of this peasant base and because of the Chinese population problem. The country as a whole could not develop at all without a developing capacity to feed its people.

Reconstruction (1949–1952) and the first five-year plan (1952–1957)

After the Communists won their political revolution in 1949, they began their economic revolution. They studied the successes and failures of the Russians and proceeded slowly to socialize the country. One of the party's revolutionary slogans had been "Land to the tillers." To cement its control over the country's producing regions (and to fulfill its revolutionary "campaign promises"), the party carried out an extensive *land reform* as one of its first major programs.

Nearly every revolutionary movement in an agricultural country promises land reform. In a narrow sense, all that this means is redistribution of ownership of land, animals, and farming implements. In itself, this is not socialism. Market incentives are left intact. Large landholdings are broken up and parceled out to formerly landless peasants. If anything, this increases the number of "capitalists," although it largely eliminates those who hire the labor power of others.

The political effect of land reform is quite important for later socialization, however. The breakup of the large holdings eliminates the political power of the landlords and wealthy peasants, and substitutes the political power of the party carrying out the reform. Thus, besides bettering the lives of the landless peasants, land reform strengthens the grip of the revolutionary party over the countryside. This makes the later collective socialization of agriculture go more smoothly than it would otherwise. In China's case, the collectivization drive of the 1950s followed soon after the land reform. In Russia, a long period intervened between land reform and collectivization. This allowed the power and prosperity of the richer peasants to grow, which may be why the Russians met with so much more resistance than the Chinese, and why Stalin's collectivization was so violent and destructive.

Collectivization in China proceeded gradually. At first, the means of production remained in private hands, and the only cooperation took the form of "mutual aid teams" that farmed the members' land in rotation. Then, during China's first five-year plan period (1953–1957), the land was pooled, but incomes were initially paid according to a formula based in part on the amount of land contributed to the pool. Only in the last stage of a seven-year process were incomes based on labor alone, effectively ending any connection between the peasants' incomes and their initial landholdings.

One of the arguments on behalf of collectivization of agriculture (rather than egalitarian land reform) is that it leads to economies of scale from combining small land parcels to permit the use of machinery that no small peasant could afford or use effectively. In China's case, these advantages seem not to have been realized. There are several reasons for this. First, it is not possible to mechanize without machinery. Since China had very little farm machinery or capacity to build it, it could not have mechanized its agriculture very fast even if there had been a big advantage in doing so. Second, mechanization involves the substitution of tractors for people. Since there was a large surplus population in the agricultural areas of China, there was no gain to releasing people from the land.

The collectivization of the 1950s did not, therefore, bring a major change in the proportions of labor relative to other inputs. Teamwork was substituted for individual cultivation, but not machinery for labor. Doubtless there were gains from this reorganization of work. Doubtless there were also losses because of the reduction in individual material incentives. Official

Chinese figures show marked gains in output throughout the period. Some Western estimates, however, show very little growth. But it is clear that there was no disruption of production in China comparable to what happened in the Soviet Union. And this is fortunate because there was very little room for error. Much of China's population was close to the margin of starvation at the beginning of the 1950s.

What did change were the ***relations of production.*** This may not strike you as an achievement, but to the Communist, the substitution of teamwork for individual cultivation was a substantial step toward the development of communism.

The reconstruction period and the first five-year plan also had an industrial and commercial dimension. Although China was far less industrialized in 1949 than Russia was in 1917, there was a substantial amount of heavy industry already controlled by the Nationalist government and an extensive privately owned industry producing consumer goods. The Communists took over the operation of heavy industry and banking almost immediately. The remainder of industry and trade were socialized gradually. At first, capitalist management was retained. After a few years, it was replaced by socialist management. But the transformation from capitalism to socialism in industry and commerce was very gradual.

Taking control of an economy and socializing it is a very complicated task. One of the greatest strengths of capitalism is the structure of profit incentives built into the price system. Producers who respond to these incentives coordinate their activities without the need for centralized decision making. When a socialist government destroys these profit incentives, it must find some other coordinating mechanism or the result will be chaos. In Russia, Lenin foresaw the difficulties he would run into if he tried to socialize the entire economy at once, with no previous experience and little technical skill. He chose to begin by taking over what he called *the commanding heights*—heavy industry, international trade, and banking. By controlling the flows of materials, capital goods, and finance capital, he hoped to control the development of the economy as a whole, even though light industry and commerce remained in private hands. The Chinese followed Lenin's example. Nationalization was carried out piecemeal rather than all at once. By limiting their immediate goals to fit their capabilities, the Chinese managed to avoid costly disruptions that they could ill afford.

The first five-year plan was similar to the Russian plans, in that it placed primary emphasis on the development of heavy industry. But unlike Stalin's strategy, the Chinese plan called neither for the oppression of the peasantry nor for the neglect of agricultural development. The resources for investment in heavy industry came from three sources. First, consumption in general was kept down, both in urban and rural areas. Second, investment in light industry and commerce was sharply restricted. Third, the Russians assisted the Chinese with substantial trade and credits. They delivered several billion dollars' worth of modern capital goods (including, for example, a complete tractor factory). The countryside was permitted to retain enough food to support its own program of improvement in agricultural productivity. The Chinese plan was more balanced than the Russian, reflecting the realities of the Chinese food situation and the importance of the peasantry in the Chinese Communist party.

But because the Chinese approach to socialization and development was balanced, moderate, and gradual by comparison to the Russian doesn't mean it was humane. Both the land reform program and collectivization involved public hu-

miliation, "reeducation," and murder of landlords and rich peasants. The so-called Three Anti and Five Anti movements, which were directed at "bourgeois elements" in administration and the urban economy, were similarly brutal. In many respects, they were hard to distinguish from the "Four Evils" campaign, which was directed at rats, flies, mosquitoes, and sparrows. But the pace of socialization was moderate enough to avoid any disastrous drop in production.

The Great Leap Forward, 1958–1960

Deployment of a large, well-equipped field army presents many of the same problems as those faced by economic planners. Achieving coordination among fighting units and providing supplies, food, and ammunition are major concerns. When the Yugoslav partisans succeeded in breaking out of the mountains late in World War II, they found that guerrilla tactics were quite ineffective against German field armies.

Maoist economic policies have often been likened to the techniques of guerrilla warfare. They emphasize spontaniety, voluntary effort, patriotism, willingness to undergo hardship, and determination to succeed in the face of material obstacles. To see what this means in concrete terms, it is helpful to examine the period known as the ***Great Leap Forward*** (1958–1960).

The Great Leap, which represented a radical departure from the economic policies of 1949–1957, had two dimensions. The first was yet another transformation of the relations of production in agriculture, another step toward the classless society. The second was a reorganization of the forces of production based on the new rural social structures.

The local political unit corresponding to the collective farm was the *cooperative*. This was a rather limited institution whose main function was to organize the operations of the farm itself and to supervise its production process. The representative cooperative consisted of 100 to 200 families. Under the Great Leap, agriculture was reorganized into far larger groups known as *communes*. A representative commune consisted of about 5,000 families. The communes took over all of the functions of the cooperatives and some of the functions of the state, such as the provision of tractor services and education. The communes also developed large-scale facilities for providing meals and child care. Thus, some activities and income were transferred from the households to the commune and made available to everyone according to need. The commune was viewed by the Maoists as a substantial advance toward full-fledged communism.

The dimension of the Great Leap that received most attention in Western countries was its program of economic development in the communal areas. The communes were directed and exhorted to embark on a massive campaign of self-development by harnessing the "revolutionary energy of the people." This energy was to be channeled in three directions:

1. Flood control and irrigation projects were to be initiated locally and carried out with local labor resources.
2. People were encouraged to invent and build their own agricultural and construction tools.
3. Small-scale industrial plants, associated with the communes, were to be built in the countryside, using locally available materials and labor as much as possible.

Both American and Russian observers viewed this program with skepticism—the Americans with amusement and the Russians with horror. The "backyard" steel furnaces, fertilizer plants, and electric

power generators were singled out for particular scorn.

In theory, the Great Leap was not such a bad idea, however. The countryside must have had an absolute surplus of labor at some times of the year, even under the cooperative. Indeed, some estimates place the surplus in the neighborhood of 30 million people! The program of water control, tool building, and creation of small-scale industry could be carried out by this surplus labor, so that the rate of development in the countryside could be increased without either diverting investment from urban industry or reducing current food production. Only a country with surplus labor on the land could raise its growth rate in such a costless way.

Unfortunately, the Great Leap seems not to have worked very well. Irrigation and flood-control dams that were built of unreinforced earth went to pieces. Handmade tools didn't work. The products of the small-scale industry were of low quality, in some cases unusable. And many peasants were not happy with the commune, which was excessively large. Traditionalist peasants preferred their families over communal eating and child-rearing arrangements.

Labor and enthusiasm may have been superabundant in the countryside, but understanding, knowledge, and the experience necessary to carry out the Great Leap projects were not.

When you were a child, you may have tried to make cookies and ended up simply making a mess of the kitchen. All the excitement and energy you could muster could not compensate for your inability to cook. In the same way, revolutionary zeal and the energy of the people cannot carry off a rural development program without the help of agronomists and engineers. If the Great Leap had been more carefully planned and the revolutionary "cadres" had been trained in basic technology as well as in Maoist political thought, it would have leapt much further.

In the early months of the Great Leap, the Chinese press reported success after success. Later on, the reports became more muted. Of particular concern were grain shortages beginning early in 1959, apparently resulting from the disruptions brought about by the Great Leap. Then, in the summer of 1959, Chinese agriculture was struck by a series of disasters—floods in some areas and droughts in others. These recurred in 1960. In no sense did they result from the Great Leap itself, but its water-control projects could not contain the problem. Grain production dropped by about 10 percent in 1959 and another 10 percent in 1960. Declines of this magnitude are common in nearly all grain-producing countries. Those countries that are relatively affluent and well integrated into the world grain market usually handle them without great problems, although bad harvests frequently slow down industrial growth in the Soviet Union. But in China, whose population is never very far from the margin of subsistence, a bad harvest is a social calamity. Malnutrition and disease were widespread in 1960 and 1961. In human terms, the bad harvests were tragic. In economic terms, the lack of food greatly weakened the ability of the affected populations to work and slowed the rate of overall development. China had to import some grain, restricting its ability to import capital goods and materials.

The failure of the Great Leap had other, more lasting, effects. First, the strength of the commune as an economic and political unit was greatly reduced. Second, the political power of the Maoists was weakened. Mao himself continued to be revered, but the Maoists lost control of economic policy. And finally, toward the end of the Great Leap, the Soviets withdrew from China their scientific and technical assistance team, and sharply cur-

tailed other forms of development assistance. This was part of a general breakdown of relations between the Russians and Chinese that has never been repaired, to this date.

China since the Great Leap

Economic policy after the Great Leap focused on the recovery and development of agriculture. This had several concrete implications. First, instead of being assigned to communes as a whole, quotas were assigned to far smaller groups within the communes. This established a much more direct link between individual performance and material rewards. Second, the cultivation of private plots was encouraged. Third, the communal eating and child-care facilities were largely abandoned. Such changes in the social relations of production represented a return to earlier ways of doing things in preference to those of the Great Leap.

The ***forces of production*** were also redeployed to assist agricultural development. The rural industrialization program was largely abandoned, but the urban industrial sector was directed to increase the production of material inputs for agriculture—especially farm machinery, pesticides, and chemical fertilizers.

This provision of material incentives and improved material inputs for the agricultural population is not a very dramatic policy compared with the Great Leap. But it seems to have worked, with some help from the weather. Agriculture recovered, and China was again able to feed its population. Most Chinese development since the Great Leap has been similarly unexciting. People have been encouraged to pursue their own immediate material interests within a framework of state ownership and planning. Planning has recognized the primacy of agriculture in the Chinese situation, and has modified the Stalinist development strategy to direct more industrial output toward agriculture.

The Maoists had one final fling, however. This was the ***Great Proletarian Cultural Revolution*** (1966–1969). It was largely a political struggle, in which the most important matter at stake was who would control the development of the economy after Mao's death. In centrally planned economies, nearly all political struggles have economic implications.

In a sense, the Cultural Revolution was Mao's reaction to the failure of the Great Leap. Mao consistently pursued the goal of a classless society. The Great Leap strategy was to attack social relations at the site of production—to reorganize work on a communal rather than an individualistic basis. Mao apparently expected cooperative relations in the workplace to replace the need for hierarchy, and for these habits of cooperation to spread throughout the network of national administration. Economic and political initiatives would rise up from self-governing communes rather than being imposed from above by a bureaucracy.

When the Great Leap failed to produce successful communes, Mao turned to a direct attack on the growing bureaucracy.

The Cultural Revolution began as a Maoist propaganda campaign directed at "bourgeois" and "bureaucratic" elements in the government and party. The Maoists urged the people to rise up against those who favored conventional Stalinist or reformist economic policies. They were painted as a new ruling class, antagonistic both to the masses and to the achievement of classless communism, like the bureaucrats of the Soviet Union who rule the country by controlling the economy. The Maoists mobilized large numbers of students and other young people into the "Red Guards," a militantly revolutionary mass movement. The Red Guards carried

out a program of public humiliation and violence against "enemies of the people," both domestic and foreign.

It is hard to know the extent to which the Cultural Revolution was a sincere attempt to prevent the government and party bureaucracy from becoming a new ruling class, as it had in the Soviet Union, and to what extent it was a sordid struggle for power. It is clear that the mass movement got out of control and damaged the lives of many officials and intellectuals who were in no sense counterrevolutionary. To an extent that is impossible to measure, it set back the development of the economy. Industrial output declined in 1967, and foreign trade was sharply curtailed for several years.

The universities were especially hard hit. University students in large numbers joined the Red Guards. Among the targets of their revolutionary fervor were their former professors, who were physically assaulted and forced to make public confessions of counterrevolutionary activities. Participating in such activities must have been fun for the Red Guards. Haven't you ever daydreamed of humiliating a hated professor? Dreams are harmless enough. But once they are acted out on a wide scale, and the offending professors are banished from teaching, there is not much left of the higher education system. A country that is very short of scientific, technical, and organizational skill can ill afford to have its higher education system closed down for a decade.

In the end, even the Maoists had to back off. The Red Guards were disbanded, and the Cultural Revolution subsided. Although it was never really called off, it effectively died out in a few years. Maoism was discredited, and moderate, more conventional, leaders regained control of economic policy. After Mao's death in 1976, the reformers were ascendant. Material incentives were strengthened, foreign trade was expanded, and individual enterprises were given progressively greater latitude in choosing what to produce and even what prices to charge. By 1980, the anti-Mao forces felt sufficiently well entrenched to put the so-called Gang of Four (including Mao's widow) on trial for the excesses of the Cultural Revolution.

At the time of the ascendancy of reform, American news media triumphantly proclaimed that China was "going capitalist." It turned out that all they meant by this was that China was giving more autonomy to individual productive units and establishing a closer connection between individual productivity and income. These features, of course, do not define capitalism. The basic traits of capitalism are *private* ownership and control of the means of production. This China does not have. The reforms simply favor greater reliance on individual material incentives, and less reliance on the revolutionary spirit so valued by the Maoists. In effect, they favor going slow on the transformation of the relations of production, in the hope of securing faster development of the forces of production.

The future of China

Winston Churchill once referred to the Soviet Union as "a riddle wrapped in a mystery inside an enigma." In later years, he may have wished he had saved this comment for the People's Republic of China. Students of Chinese economic development have seen so many twists and turns of strategy during their lifetimes that they are very cautious about making predictions. Partly this stems from lack of solid information. Chinese society is so vast and undeveloped that the Chinese themselves probably don't understand it very well, and they are quite secretive about what they do know. But partly it stems from an unresolved conflict between those who

give primacy to development of the forces of production (economic development) and those who give primacy to revolutionizing the relations of production (social development). You should not be surprised if this conflict breaks out in another form during the 1980s.

Summary

This chapter has developed a lot of themes, some of them quite general, and some specific either to the Soviet Union or to China. Among the more important are the following:

1. Economic life in much of the world is organized along socialist rather than capitalist lines.
2. The major socialist revolutions have taken place in economically backward countries, so the major problem faced by socialist governments has been that of economic development.
3. The Soviet Union, under Stalin, stressed development through forced industrialization. This involved squeezing the agricultural sector to provide resources for the development of heavy industry. The mechanism for extracting a surplus from agriculture was the collective farm. It consolidated land and labor into a large organization that was forced to trade on unfavorable terms with the state.
4. The Soviet development strategy transformed the Russian economy. Today, Russia is a modern industrial society. Its economic life is organized by a planning structure rather than by a network of markets. Russian planning allows a degree of choice to households, but almost none to enterprises, which are told what and how to produce.
5. The Soviet planning structure is not very efficient in its use of existing resources, but it has historically produced rapid rates of growth.
6. The Russian economy is not very egalitarian, since it uses bonuses, wage differentials, and special privileges to reward those who are thought to be especially productive. The bureaucracy is very large and powerful. In some respects, it resembles the ruling class of capitalist societies.
7. The development strategy followed by the Chinese has differed from that of the Russians by emphasizing agricultural development as well as industrialization. The main reason for this is the Chinese population problem.
8. Since the Chinese Revolution, periods of stability and economic growth have alternated with episodes of great disruption (the Great Leap Forward and the Proletarian Cultural Revolution).

Key concepts

Centrally planned economy
Forced industrialization
Collectivization of agriculture
Material incentives
New ruling class
Relations of production
Great Leap Forward
Forces of production
Great Proletarian Cultural Revolution

Questions for review

1. Why must forced industrialization come at the expense of the agricultural sector?
2. In what sense does the planning bu-

reaucracy in the Soviet Union serve the same functions as the price system in a capitalist country?

3. Why is central planning an almost unmanageable process?
4. How did China's population problem influence its development strategy so as to make it different from the Russian strategy?
5. To what extent do you think people are guided by their own immediate economic self-interest? How does this influence your attitude toward Maoist economic policies?

Glossary

absolute advantage (34) See *comparative advantage.*

accounting cost (8) See *direct cost.*

accounting profit (7) See *profit.*

accumulation of capital (37) The conversion of surplus value into material inputs of production, such as machinery, factories, and equipment.

allocative efficiency (9) Is achieved when price equals marginal cost for every firm.

antitrust (12) Policies that deal with anticompetitive conditions that arise in various markets.

arbitrage (35) Involves the simultaneous buying and selling of a currency at two different prices. If the dollar is cheaper relative to the Deutsch mark in London than it is in Hamburg, arbitragers will simultaneously buy dollars in London and sell them in Hamburg.

assets (7) Assets include two categories: current assets are mainly cash, accounts receivable, and inventories; fixed assets are the real plant and equipment that the firm has built up over the years.

autonomous demand (25) That portion of aggregate demand that does not vary with income.

average fixed cost (8) See *fixed cost.*

average product (8) See *total product.*

balance of payments The balance between demand for and supply of a country's currency on the world's currency markets.

balance sheet (7) A financial statement that lists the firm's assets and liabilities (or claims against its assets).

bank reserves (28) Vault cash and interbank deposits. Banks who are members of the Federal Reserve System must hold their reserve deposits at the Fed. A nonmember institution must hold its reserve deposits at the Fed, at a bank which has an account at the Fed, or at one of the agencies that regulates thrift institutions.

bilateral monopoly (15) A market in which both buyers and suppliers have market power. In labor markets, a bilateral monopoly would be a monopsonist (demand side) facing a union (supply side).

bourgeoisie (37) See *capitalist class.*

bracket creep (18) When the share of personal income taken by the government in taxes increases because of increases in nominal (but not real) income.

built-in stabilizers (31) A feature of the economy which automatically works to dampen fluctuations in income. Examples of built-in stabilizers include income taxes and unemployment compensation.

business cycle (22) Year-to-year fluctuations around the steady growth trend of GNP. The business cycle is marked by peaks of prosperity and troughs of depression or recession.

business sector (23) The sector of the economy that produces goods and services and buys input services. The business sector also buys both investment and intermediate goods produced within the business sector itself.

capital (16) Human-made inputs to the process of production. Physical capital includes such tangible capital as buildings, machinery, roads, and telephone systems. Human capital refers to the skills and talents of the labor force. Money capital refers to assets such as stocks and bonds which are sold to finance the purchase of physical capital.

capital account (35) See *current account.*

capitalism Refers to economic systems in which capital is mainly privately owned rather than owned by the government.

capitalist class, bourgeoisie (37) The owners of property and the top managers who provide material inputs, decide what to produce and how to produce, and control the final distribution of the product.

cartel (34) An association of producers which sets prices and enforces penalties against members who violate the agreement.

checking deposits (27) See *demand deposits.*

circular flow (2) Interconnections of selling, consuming, and producing which tie the sectors of the economy together, so that events in one sector affect the economic system as a whole. The real flows of goods and services are matched by money flows moving in the opposite direction.

classical economics A school of economic thought characterized by its insistence that national wealth is the capacity (resources, labor, and capital) to produce goods or material products.

closed shop (15) See *labor union.*

commercial banks (27, 28) Financial intermediaries that accept deposits from households, firms, governments, foreigners, and other financial institutions. They also make loans and buy negotiable debts.

comparative advantage (34) Countries have a comparative advantage in the production of a good when they can produce it relatively more efficiently than other goods, when compared to other countries.

Countries have an absolute advantage when they can produce a good more efficiently than another country.

competition (9) Pure competition is an industry model based on the assumptions that: a) there is one identical good sold in the market; b) no firm has a significant market share; c) firms adjust quickly to changes; and d) there is freedom of entry and exit. Perfect competition adds the assumption that firms and consumers know prices throughout the market.

concentration ratio (10) The combined share of a market's sales that is made by the largest firms in the industry. Concentration ratios are reported for the top 4, 8, 12, and 20 firms in each industry.

conglomerate firm (11) A firm which operates in more than one market. Conglomerates range from firms with two product lines up to highly diversified enterprises with hundreds of branches and thousands of products.

conglomerate merger (11) See *merger.*

constant dollar GNP (23) See *gross national product.*

consumer price index (CPI) (23) See *price indexes.*

consumers' surplus (10) The difference between the value which the consumer places on a good and the price which the consumer must pay for the good.

consumption function (24) The relationship between consumption (C) and disposable income (YD) is called the consumption function.

consumption goods (23) The outputs of the business sector that go to households.

cooperative (18) An enterprise owned by its customers or suppliers, with profits channeled back to its owners.

corporation (7) A firm that issues voting stock which investors can buy and sell.

cost-benefit analysis (18, 20) A method of allocating scarce resources efficiently by equating marginal social costs and benefits.

cost-push inflation (26) See *inflation.*

Council of Economic Advisers (CEO) (30) A three-member group, appointed by the president with the consent of the Senate, that is responsible for advising the president on economic policies.

cross-elasticity of demand (4) $\frac{\%\Delta Q \text{ of Good 1}}{\%\Delta P \text{ of Good 2}}$ Measures how the quantity demanded of one good responds to price changes of another good. A positive cross-elasticity indicates that the goods are substitutes. A negative cross-elasticity indicates that the goods are complements. Complementary goods are those used together such as cars and gasoline.

crowding out (29, 30) Occurs when an increase in government spending causes interest rates to rise, reducing or crowding out private sectors spending.

current account/capital account official settlements (35) The demand for and supply of dollars are grouped into three categories: (1) Current account—largely imports and exports, plus small amounts of transfer payments (such as foreign aid, pensions paid to Americans living abroad, and remittances sent home by immigrant workers); (2) Capital account-lending and investment across national boundaries, some of it by individuals, some by multinational corporations; (3) Official settlements-transactions that are government interventions in the currency markets to stabilize currency values.

demand (4) The entire price-quantity relationship; the entire demand curve.

demand deposits (27) The debtor (i.e., the bank) must pay off the debt on demand either to the depositor or to anyone else whom he or she designates by writing a check. Such accounts are often called checking accounts.

demand-pull inflation (26) See *inflation.*

dependent variable (3) A variable which is influenced by changes in another variable. In an equation relating expenditures and income, expenditures is the dependent variable, being influenced by changes in income.

depreciation (23) The rate at which capital goods are being lost through wear, breakage, and obsolescence.

depreciation allowances (23) See *gross business saving.*

diminishing marginal effect (2) The effect of any good or input tapers off as more of it is used.

diminishing marginal returns (law of) (2) States that when additional units of one input are used, with other inputs held constant, a point is reached beyond which marginal product begins to decline.

diminishing marginal utility (6) As the amount of a good consumed by an individual increases, marginal utility falls.

direct cost/explicit cost/accounting cost (8) Direct, explicit, and accounting cost are the same: that part of a firm's cost that can be calculated by adding up the dollars the firm pays out. Direct costs include the purchase of raw materials and equipment, wages paid to hired employees, rent, interest, and utilities.

discount rate (31) The rate that the Federal Reserve charges those banks that borrow reserves from the Federal Reserve.

diseconomies of scale (8) See *economies of scale.*

disposable income (23) See *personal income.*

distributions (3) Graphs that show how such economic variables as income and wealth are spread among the population.

disutility (6) Displeasure, as indicated by negative marginal utility.

dividends (7) Payment of some of the accounting profits to shareholders by the firm.

dominant firm (10) One firm has 40-100% of the market and no close rival.

dual economies (36) See *uneven development.*

economic analysis (1) A system of concepts and logical hypotheses developed over more than two centuries in debates among economists.

economic cost (8) A firm's economic cost is the opportunity cost of production. Economic cost is the sum of all direct and imputed (implicit) costs, thereby including a value for all scarce resources used in production.

economic model (3) A precise formal statement of one or more economic relationships.

economic profit (8) See *profit.*

economic rent (14) A payment to an input in excess of the payment necessary to elicit supply.

economic system (1) Each economy is a system in which the production and distribution of goods are organized around society's wants. Modern economic systems range from freemarket capitalism, in which most choices are made in private markets, to controlled economies in which choices are determined primarily through economy-wide plans.

economies of scale/diseconomies of scale (8) Refers to the decrease in long-run average total cost which can accompany increases in output. Economies of scale result from specialization, physical laws, and management.

Increases in long-run average total cost accompanying increases in output are referred to as diseconomies of scale.

Economies and diseconomies of scale explain the U-shape of the long-run average total cost curve.

efficiency (2) Efficiency in the use of resources is achieved when a given level of output is produced using the least amount of inputs. Economic efficiency in production is referred to as the least-cost method of production.

elasticity (4) A measure of the responsiveness of one variable to a change in another variable.

elasticity of demand (4) See *price elasticity of demand.*

elasticity of supply (4) Elasticity of supply: $\frac{\%\Delta Q \text{ supplied}}{\%\Delta \text{Price}}$ measures the relative responsiveness of quantity supplied to changes in price. Because price and quantity supplied are directly related, supply elasticity will be positive. If elasticity is greater than 1, supply is said to be elastic, with quantity supplied being relatively responsive to changes in price. If elasticity is less than 1, supply is said to be inelastic, with quantity supplied being relatively unresponsive to changes in price. In most cases, supply elasticity will differ from one point on the supply curve to another.

enterprise (7) Enterprises or firms are the basic units of production, converting inputs into outputs. An enterprise may consist of one local plant (factory or office) or up to hundreds of plants.

equilibrium (2) A condition reached when all influences balance one another out, so that there is no pressure for further change.

equities (27) See *loans/negotiable debt/equities.*

equity (7) The net worth of the firm, equal to assets minus liabilities.

excess reserves (28) See *reserve requirements.*

exchange rate (35) The rate at which one currency may be exchanged for another. In the case of fixed exchange rates, a government is committed to buying and selling its own currency in order to maintain the agreed upon value of the currency. In the case of fluctuating exchange rates, the value of a currency is allowed to change as supply and demand conditions change.

expectational inflation (26) See *inflation.*

explicit cost (8) See *direct cost.*

export promotion/import substitution (36) Export promotion, an outward-looking policy, refers to the development of export industries. Import substitution, an inward-looking policy of diversification, refers to the development of domestic sources of supply for goods previously imported.

exports (23) Goods and services produced in one country and sold to another country.

external effects: costs and benefits (17) External effects—either costs or benefits—occur whenever an action has repercussions which the actor need not take into account. In such cases, social costs and benefits (costs and benefits to the entire society) will differ from private costs and benefits.

factor income/wages/property income (23) Income paid to owners of inputs or factors of production. Wages are payment for labor services. Property income is from rent, interest, or profit. Wages and property income make up the factor income which gives households the ability to buy goods and services.

factors of production (2) See *inputs.*

Federal Reserve (Fed) (28) The agency responsible for regulating banking in the U. S. and carrying out monetary policy.

financial intermediaries (27) Banks, savings institutions, pension funds, life insurance companies, and the like, through which funds flow from savers to borrowers.

financial markets (27) Financial markets complete the circular flow by moving

funds from the surplus units to units that wish to run deficits.

fixed cost/average fixed costs (8) Costs of production which do not vary with the level of output. Examples of fixed costs include rent and bank payments. Fixed costs exist only in the short run. Average fixed cost refers to fixed cost per unit of output.

fixed exchange rate (35) See *exchange rate.*

flow (3) Processes or values occuring over a period of time. Income per year or sales per month are examples of flows.

forced industrialization (38) A program to induce rapid growth in the industrial sector relative to other sectors of the economy.

forces of production/relations of production (37, 38) Forces of production are labor power, material inputs, and technology. Relations of production are the social setting of production.

foreign sector (23) The foreign sector supplies imports and buys exports.

GNP deflator (23) See *price indexes.*

GNP effect (29) One of two paths linking the financial and goods markets. The GNP effect runs from GNP to the rate of interest by way of the money demand curve. A rise in GNP shifts the demand curve for money to the right and raises the rate of interest.

GNP gap (30) See *potential GNP.*

general equilibrium (17) Refers to equilibrium of the entire economy. General equilibrium analysis examines the impact of a change in one sector of the economy on the economic system as a whole.

government sector (23) The government sector of the economy buys factor services and goods from the private sectors. It provides them with government services, which are largely financed through taxation. The output of the government sector includes defense, education, fire and police protection.

gross business saving/net business saving (23) Gross business saving is equal to the sum of retained earnings and depreciation allowances. Retained earnings are profits which are not paid out in dividends. Depreciation allowances are funds kept by firms to replace machinery and equipment which has worn out or depreciated with use.

Net business saving equals gross business saving minus depreciation. It therefore equals retained earnings.

gross fixed investment/net fixed investment (23) The total of fixed capital goods (machinery, factories, homes) produced in a given year. Net fixed investment equals gross fixed investment minus depreciation.

gross national income (23) The total payment to inputs or factors of production within a given period of time. It is equal to the sum of wages, property incomes, and indirect taxes paid by business.

gross national product (GNP)/net national product (NNP) (22,23) The market value of all goods and services produced in the economy in a given period of time. It is equal to the sum of consumption investment, government purchases, and net exports. Net National Product is equal to Gross National Product minus depreciation.

high-employment budget (31) A tool developed by the CEA to chart changes in fiscal policy. It keeps the budgetary changes that result from the business cycle separate from those that result from fiscal policy changes. The high-employment budget calculates receipts and expenditures that would result from existing legislation if the economy were operating at its potential GNP.

high-powered money (28) See *monetary base.*

horizontal merger (10) See *merger.*

household (2) Any unit in which people make decisions about work, consumption,

the disposal of personal property, and other personal activities. Households consume, as opposed to enterprises, which produce.

household sector (23) The household sector of the economy supplies factor services and buys consumer goods.

human capital (15) See *capital.*

immigration (33) See *natural increase.*

implicit cost (8) See *imputed cost.*

import substitution (36) See *export promotion.*

imports (23) Imports are goods which are produced in other countries and brought into one country.

imputed cost/implicit cost (8) The estimated or implicit value of scarce inputs for which there is no market transaction. Their value is equal to the return they would get in their best or highest paying alternative use.

incidence (18) See *tax incidence.*

income effect (4,15) As the price of a good changes, purchasing power changes, so consumers must adjust the quantity demanded of all goods. In labor markets, the income effect refers to the increased ability of workers to purchase leisure as the wage rate increases.

income elasticity of demand: normal good and inferior good Income elasticity measures the responsiveness of quantity demanded to changes in income. A positive income elasticity indicates that the good is a normal good, with quantity purchased increasing as income increases. A negative income elasticity indicates that the good is an inferior good, with quantity purchased decreasing as income increases. If the value of income elasticity is greater than 1, the good is said to be a normal good with an income-elastic demand, meaning the quantity demanded is relatively responsive to changes in income. If the elasticity value is between 1 and zero, the good is said to be a normal good with an income-elastic demand, meaning the quantity demanded is relatively unresponsive to changes in income. Income elasticity can vary with income levels.

income statement (7) A statement showing the firm's revenue and the division of its revenue into costs, taxes, dividends, and retained earnings.

independent variable (3) A variable which causes change in another variable. In an equation relating expenditures and income, income is the independent variable, influencing the level of expenditures.

indexing (30) A way to protect people against the redistributive costs of inflation. Tying all contractual payments to some price index such as the CPI.

induced demand (25) The portion of aggregate demand which varies with income. Induced changes in demand are caused by changes in income.

industry (7) A set of producers making similar products.

inferior good (4) See *income elasticity of demand.*

inflation An increase in the general level of prices. Inflation has many causes. Cost-push inflation results from increases in input costs. Demand-pull inflation results from an expansion of demand. Expectational inflation results from the fact that when people expect inflation, their behavior (such as negotiating a wage increase or a long-term contract) is altered in ways which will cause the inflation rate to increase.

innovation (16) See *invention/innovation/imitation.*

input-output tables (23) Detailed studies, published by the Department of Commerce, of the flow of goods from one industry or sector of economy to another. They form the basis for estimates of value added by the various industries.

Inputs (or factors of production) (2) Items used in the process of production. The three traditional classes of inputs are labor (physical and mental effort), capital (human-made aids to production), and land (natural resources).

Intercept (3) The point at which a straight line touches the vertical axis.

Interest effect (29) One of two paths that link the financial and goods markets. The interest effect runs from the interest rate to GNP by way of the planned demand schedule. A rise in the interest rate shifts the planned demand schedule down, lowering equilibrium GNP.

Interest rate (29) The rate of return for a lender and the cost of credit for a borrower.

Intermediate good (23) A good which is produced by one firm and sold to another firm as an input.

Internal migration (33) See *natural increase.*

Internal rate of return (16) The interest rate that will just equate the capitalized value of an investment's future profits with the investment's initial cost.

International monetary fund (IMF) (35) An international lending organization established in 1945 to "bail out" countries whose reserves have reached critically low levels.

Invention/innovation/imitation (16) Technical change can be divided into three phases: (1) invention-the creation of a new idea; (2) innovation-making practical use of the idea; and (3) imitation-diffusion of the idea as it is copied by others.

Inventory investment/inventory stocks (23) Inventory investment is one of three forms of investment: net additions to business inventories. Inventory investment can be positive or negative, depending on whether the total stock of inventories is growing or declining.

Investment (23) Investment consists of three components: (1) Capital goods, such as machinery, buildings, equipment; (2) Changes in the level of inventories (positive or negative); (3) Net additions to the housing stock. Gross investment equals the total of all investment. Net investment equals gross investment minus depreciation.

Invisible hand (17) A phrase coined by Adam Smith, who believed that the guiding of economic choices by self-interest would lead to the right amount and mix of output.

Keynesian economics (22,29) Stresses the need for government intervention to stabilize the economy.

Labor force participation rate (15) The percent of each group of the population that works for pay.

Labor monopolies (15) See *labor union.*

Labor union (15) An association of workers formed to gain some control over the supply side of the labor market.

In a closed shop, only union members may be hired. In a union shop, nonunion members may be hired if they join the union within a specified time. In an open shop, union membership is not required for employment.

Linear equation (2) The equation for a straight line representing the relationship between two variables. The general form of the equation is $y = a + bx$, where y is the dependent variable, x is the independent variable, a is the intercept of the line and b is the slope.

Loans/negotiable debts/equities (27) Loans (in legal terms, "notes") are promises to repay principal and interest according to a definite schedule. Negotiable debts are also promises to repay on a definite schedule, but unlike notes, they may be sold by the initial purchaser to someone else—i.e., they are negotiable. The buyer gets the

right to receive the payments promised thereafter. Negotiable debts include bonds, some mortgages, and a variety of other promises to pay. Equities are promises to pay to their buyers a share of the issuer's profits in the form of dividends, when and if such dividends are declared. The best-known kinds of equities are corporate stocks, which are negotiable.

long run (8) A period of time long enough for all variables to be altered in quantity.

M1 and M2 (27) Two definitions of the U.S. money supply. M1 is made up of currency (coins and paper money) and all demand deposits in banks and savings institutions. M2 equals M1 plus funds held in savings accounts, small time deposits, money market funds, and other short-term deposits of various kinds.

macroeconomics Deals with issues involving the entire economic system, such as the general level of prices and the total amount of output and employment.

Manchester School (22) Justified industrial exploitation by claiming it was necessary to achieve efficiency in free markets.

marginal analysis Most economic choices involve marginal changes. Marginal means (to be most precise) adding just one more unit. A decision "at the margin" compares the benefits and costs of changing a level slightly.

marginal cost (13) The change in total cost resulting from a one-unit increase in output.

marginal cost pricing (13) The setting of price equal to marginal cost at a level of output corresponding to minimum average total cost. Marginal cost pricing is one criterion which regulatory agencies could use to set prices of regulated firms.

marginal demand propensity (MDP) (25) The change in planned demand resulting from a change in national income or GNP (planned demand/income). The marginal demand propensity is the slope of the planned demand function.

marginal product (8) The change in output resulting from a one-unit increase in the quantity of the variable input.

marginal product/price ratio (8) Represents the increase in output resulting from the last dollar spent on the input.

marginal propensity to consume (MPC) (24) Measures the change in consumption that will result from a change in income (consumption/income). The marginal propensity to consume is the slope of the consumption function.

marginal propensity to save (MPS) (24) Measures the change in saving that will result from a change in income (saving/income). The marginal propensity to save is the slope of the saving function.

marginal revenue (9) The change in revenue resulting from a one-unit change in output.

marginal revenue product (14) The increase in revenue resulting from the use of the last unit of input. It equals the marginal product of the input multiplied by the marginal revenue of output.

Value of Marginal Product is the term used to refer to the marginal revenue product of a perfectly competitive firm.

marginal utility (6) The change in total utility or satisfaction resulting from a one-unit change in the consumption of a good.

marginal utility/price ratio (6) The change in total utility or satisfaction resulting from a one-dollar change in the amount spent on a good.

market (2,4) A grouping of buyers and sellers who exchange a specific good at a particular price.

market equilibrium (5) The price at which the quantity supplied and quantity demanded of a good are equal.

market failure (17) Markets fail to reach optimal conditions because of the interfer-

ence of such conditions as social goods, external effects, inequitable distribution, common-property resources, and monopoly.

market power/monopoly power (10) A firm with market power (monopoly power) has some degree of control over prices. This is indicated by a downward-sloping demand schedule for the firm. Firms with market power range from monopolistic competitors who have a slight degree of market power to pure monopolists who have 100% of the sales in a market.

market share (10) A firm's share of the market is measured by its own sales, taken as a percentage of all sales in the market.

Marxists (22) Followers of Karl Marx's (1818-1883) theories of economics. Marxists view capitalism as an intrinsically self-destructive social organization that will one day disintegrate from its own contradictory tendencies.

material incentives (38) Refers to a feature of planned economies in which planners arrange different wage scales to assure labor mobility toward those sectors whose growth is planned to be higher than average. Within enterprises, there is an elaborate system of wage differentials and bonuses to reward good performance.

medium of exchange (27) A primary function of money is to be used in exchange for goods and services. In the U.S. economy, both currency and demand deposits function as a medium of exchange.

mercantilism (22) The belief that economic wealth was embodied mainly in precious metals.

merger (10) The joining of two or more separate firms into a combined firm. The three main kinds of mergers are: (1) Horizontal merger: a merger of firms in the same market; (2) Vertical merger: a merger of firms at different but related steps in the production chain; (3) Conglomerate merger: a merger of firms which do not operate in related markets.

microeconomics The branch of economics that concentrates on the details of the economy, on parts of the whole.

minimal optimal scale (minimal efficient scale) (8) The lowest level of output necessary to achieve minimum average total cost.

minimum wage (5) The lowest wage which is legally permissible. The minimum wage is a form of price floor.

Monetarists (30) Economists who believe that the economy has inherent self-stabilizing properties. Monetarists tend to believe that fiscal policy is ineffective in the long run, while monetary policy is powerful but destabilizing.

monetary base/high-powered money (28, 31) The sum of the outstanding currency and net reserve deposits on the Federal Reserve balance sheet.

monetary feedback (29) The changes that make up the interest effect and the GNP effect form a closed path from changes in the interest rate to changes in GNP and back to changes in interest. Because this closed path is similar to the "feedback loops" that engineers use to stabilize mechanical and electronic equipment, it is often called the monetary feedback.

monetary policy (29) Deliberate attempts to stabilize the economy through changes in the money supply, which lead to changes in the interest rate and therefore in GNP.

money (27) A universal medium of exchange; the power to command marketable goods and services; universal equivalent or unit of account; store of wealth.

money capital (16) See *capital.*

monopolistic competition (10) A market structure in which there is a low level of concentration, free entry, and product dif-

ferentiation which gives each firm a slight degree of monopoly power.

monopoly (10) In pure monopoly, one firm has 100% of the market. A degree of monopoly or market power exists if the firm's demand schedule slopes downward.

In the case of a natural monopoly, scale economics permit the efficient operation of only one firm in the market.

monopsony (11) Occurs when a firm purchases or hires such a large percentage of inputs in a given market that the firm becomes a monopoly on the buying side of the market.

multilateral trade (34) Trade involving many countries.

multiplier (25) The process by which an autonomous change in demand is multiplied into an even larger increase in equilibrium income. The size of the multiplier is equal to 1/(1 - Marginal Demand Propensity). If the multiplier number equals 2, every $1 autonomous change in planned demand will lead to a $2 change in equilibrium income.

national debt (31) The cumulative total of the deficits in the federal budget.

natural increase/immigration/internal migration (33) Natural increase is the excess of births over deaths. Immigration is the influx of people from other countries. Internal Migration is the flow of people from one section of the country to another.

natural monopoly (13) See *monopoly.*

natural unemployment rate (26) The level of unemployment which corresponds to the level of output at which there is no demand-pull inflation. It is the lowest unemployment rate which is compatible with stable prices.

negotiable debts (27) See *loans.*

neoclassical economics (22) Focused greater attention on consumer demand and mathematical techniques.

net business saving (23) See *gross business saving.*

net fixed investment (23) See *gross fixed investment.*

net national product (23) See *gross national product.*

neutrality of money (29) The belief that the money supply determines only the price level and the level of money wages. Real variables such as real GNP, the real interest rate, and the real wage rate are independent of the nominal money supply. The nominal money supply therefore has a neutral role in determining the real economic variables.

nonprofit firm (10) A firm "owned" by a charitable group which has a special social purpose. Examples include most hospitals, the Red Cross, and city orchestras.

normal good (4) See *income elasticity of demand.*

normal return (8) Corresponds to an economic profit of zero. It represents the rate of return equal to the return the firm could make in the most profitable alternative use of resources.

normative economics (1) Normative economics concerns value judgments, or what ought to be. A normative statement usually expresses ethical standards and values.

oligopoly (11) A market structure dominated by a group of leading firms. There is often a fringe of smaller competitors. A distinctive characteristic of oligopoly is the recognition of interdependence by the firms in the industry.

Tight oligopoly refers to markets in which the four largest firms account for 60 percent of market sales. Loose oligopoly refers to cases in which the four largest firms account for 20 to 40 percent of market sales.

OPEC (26) The Organization of Petroleum Exporting Countries, which formed a

cartel (or agreement) in 1973 to set the prices at which they would sell crude oil and to divide up the market among themselves.

open market operations (28,31) One of three major tools the Fed uses to regulate the rate of growth of the money supply. Open market operations refer to the Fed's purchase or sale of government securities on the open market in order to create and destroy reserves.

open shop (15) See *labor union.*

opportunity cost (8) The cost of taking one action in terms of foregone alternatives.

perfect competition (9) See *competition.*

persistent inflation (26) An inflation that is widespread and long-lasting enough to become the expected state of affairs.

personal income/disposable income (23) Personal income refers to the portion of national income which is received by households. It is the sum of the wages and property income received by households plus government transfer payments to households. Disposable income equals personal income minus personal taxes.

Phillips curve (26) Named for the British economist A. W. Phillips, who studied wages and unemployment in late 19th and early 20th century Britain. He found that if he graphed the yearly percentage change in money wages rates against the unemployment rate for the same year, his data described a curved, inverse relationship.

physical capital (16) See *capital.*

physiocrats (22) Economists who felt that wealth was productive capacity, primarily the fertility of land.

planned demand function (25) A schedule showing what planned aggregate demand will be at every level of GNP.

positive economics (1) Positive economics concerns facts: What is happening or has happened.

potential GNP (potential output) (30) The maximum value of all goods and services that the economy can produce without generating shortages and widespread inflation—in other words, without straining its capacity. Potential GNP minus actual GNP equals the GNP gap. A positive gap implies that the economy could produce more than it is, and unemployment exists. A negative gap implies that the economy is operating beyond its capacity, and that there are inflationary pressures.

price ceiling (5) See *price controls.*

price controls (5) Legal maximum or minimum prices. Price ceilings set the maximum price allowed by law, while price floors set the minimum price allowed by law.

price discrimination (10) Setting different price-cost ratios for different customers, rather than one price/cost ratio for all.

price effect (15) The increasing cost of leisure resulting from higher wages.

price elasticity of demand (elastic demand, inelastic demand, unitary demand (4) Price elasticity of demand, $\frac{\%\Delta q}{\%\Delta P}$, measures the relative responsiveness of quantity demanded to changes in price. Because of the inverse relation between changes in price and quantity, price elasticity will always be negative. The tradition is to ignore the negative sign. If the absolute value of the price elasticity is greater than 1, demand is elastic, or relatively responsive to changes in price. If the absolute value of price elasticity is less than 1, demand is inelastic, or relatively unresponsive to changes in price. When price elasticity equals 1, demand is called unitary elastic. In most cases, elasticity will differ from one point on the demand curve to another.

price floor (5) See *price controls.*

price indexes (CPI, PPI, GNP deflator)

(23) The U.S. Bureau of Labor Statistics has developed two price indexes to study changes in prices. The CPI (consumer price index) measures the retail cost of a typical urban family's consumption bundle, including goods and services, interest rates, and property taxes. The PPI (producer price index) focuses on interindustry transactions at the intermediate stages of production—raw materials, semifinished goods, and finished goods without their retail markups. The PPI is an important indicator of future consumer prices. The GNP deflator changes GNP from current dollars to constant dollars (measured in the prices of a given year, called the base year) in order to study variations in real GNP over time.

private enterprise (7) A firm owned by individuals or by other firms and operated with the primary aim of making profits for its owners.

producer price index (PPI) (23) See *price indexes.*

production efficiency (9) Average cost is as low as possible for a given level of output.

production function (8) Shows the relation between inputs and output. It is often stated as a mathematical equation.

production-possibility boundary (2) A downward-sloping boundary showing the maximum amount of goods which an economy can produce if all of its resources are used efficiently.

productivity (16) The amount of output resulting from the use of inputs. Productivity can be stated as output per unit of input (Average Product) or the increased output resulting from an additional unit of input (Marginal Product).

profit (accounting and economic) (7) Accounting profit refers to a firm's total revenues minus dollars paid out. It is equal to revenue minus direct costs. Economic profit refers to a firm's total revenue minus economic cost. Economic cost is the sum of direct costs (dollars paid out) plus imputed or implicit costs.

progressive tax (18) See *tax.*

proletariat (37) See *working class.*

property income (23) See *factor income.*

protectionism (34) A deliberate government policy of helping domestic industries meet import competition.

public enterprise (13) A firm, such as the U.S. Postal Service or the Tennessee Valley Authority, which is owned by the citizens through their government.

public good (social good) (18) Consumption of the good by one person does not reduce the supply available to others. Examples of such goods include parks, roads, and fire protection.

quantity control (5) A legal restriction on the quantity of a good or service that a producer may supply.

quantity demanded (4) A particular point on the demand curve, representing a specific price-quantity combination.

quantity supplied (4) A particular point on the supply curve representing a specific price-quantity combination.

quotas (34) See *tariff quota.*

rate of return (29) A ratio equal to the annual income returned to the owner divided by the amount of investment.

rational expectations (30) The belief that the public correctly assesses the implications of current events and policies. If the government announces a policy which is capable of stopping inflation, inflationary expectations will immediately cease. An opposing view is that inflationary expectations will only cease when the actual inflation rate begins to decline.

real money supply (money supply/price level) (29) Money expressed in dollars of constant purchasing power.

real rate of interest (29) Rate of interest expressed in dollars of constant purchasing power. This is the interest rate relevant for investment decisions.

regressive tax (18) See *tax.*

relations of production (37,38) See *forces of production.*

reserve army of the unemployed (37) Marxist term for capitalism's unemployed population.

reserves (28) See *bank reserves.*

reserve requirements (28,31) The legally required amount of reserves that a bank must hold. Reserve requirements are set by the Fed and can be used as a tool of monetary policy. Excess reserves are reserves which a bank is holding beyond the legally required amount. Excess reserves equal actual reserves minus required reserves.

retained earnings (23) See *gross business saving.*

revenue (7) Total revenue is the price of each unit of output times the quantity of output sold.

sales tax (5) A common form of governmental levy, imposed in most states of the U. S. It is a tax levied on goods when they are sold. An example of a sales tax would be a per gallon tax on each gallon of gasoline sold, or a tax per package of cigarettes. The consumers' burden of a sales tax is the increase in price per unit of the good which results from the tax. The producers' burden of a sales tax is the reduction in revenue per unit of the good which results from the tax.

savings function (24) The relationship between savings (S) and disposable income (Yd).

savings or thrift institutions (27) Include savings and loan associations, mutual savings banks, and credit unions. All accept small deposits, pay interest or dividends, and allow their depositors to withdraw money on fairly short notice without a lot of red tape. These institutions make consumer loans and deal in the securities and mortgage markets, accumulating diverse assets.

security exchanges (27) Markets on which negotiable equities and debts (stocks and bonds) are bought and sold.

short run (8) A period of time during which at least one input is fixed.

slope (3) The degree of slant or tilt of a line.

social benefits and costs (20) See *external effects.*

social good (18) See *public good.*

specialization (8,34) Performing a task repeatedly so that one can become fast, skilled, and efficient at it.

In trade theory, efficiency in world production is reached when all countries specialize in the production of those goods for which they have a comparative advantage.

stabilization policy (30) Deliberate government action to smooth out the business cycle and achieve both full employment and stable prices.

Stalinism (38) Stalin's economic development program, launched when he gained power in Russia in 1928. It included forced industrialization and collectivization of agriculture.

stock (3) A value at a point in time, such as wealth held by an individual.

substitution effect (4) The change in quantity demanded of a good that occurs as consumers substitute one good for another in response to a price change.

supply (4) The entire relationship between price and quantity supplied; the entire supply schedule.

supply-side economics (32) Emphasizes the need for policies to increase productivity of the nation's inputs, in order to solve

the problems of unemployment and inflation.

surplus value (37) Labor produces commodities. Some commodities go back into the production process to replenish material inputs used up in production. Another portion of the commodities goes to the work force as wages in order to reproduce labor power. What is left over is surplus value: The collection of commodities destined for other uses than reproduction of human and material inputs.

tariff quota (34) Tariffs and quotas function as restraints on trade. A tariff is a tax on imports that raises the prices of foreign goods relative to domestic substitutes. A quota is a legal restriction on the quantity of a good that may be imported.

tax (18) The taking of money from a person or organization by a government. There are personal taxes such as income taxes and *in rem* taxes (taxes on things) such as sales taxes and property taxes.

A progressive tax is any type of tax which takes a higher percentage of income as income rises, falling more heavily on the rich. A regressive tax takes a decreased percentage of income as income rises, falling more heavily on the poor. A proportional tax takes the same percentage of income at every level of income.

tax friction (18) The loss of production that occurs when people alter their economic decisions in order to lighten their tax burdens.

tax incidence (18) Analysis of who bears the burden of a tax.

technology (8) The state of the art with respect to production. It encompasses all of the known methods of production.

terms of trade (34) The prices of a country's exports relative to the prices of its imports.

third world (36) The underdeveloped or least developed or least industrialized countries in the world.

thrift institutions (27) See *savings or thrift institutions.*

time and savings deposits (27) Time deposits mature at a definite date and entail an interest penalty if they are withdrawn earlier. Savings deposits pay interest steadily and may be withdrawn at any time or with short notice.

time series (3) A graph that shows how an economic variable has moved or behaved over time.

total cost/average total cost/marginal cost (8) Total cost refers to the total cost of production; the sum of fixed and variable costs. Average total cost is total cost per unit of output. Marginal cost is the change in total cost resulting from a one-unit change in output.

total product/average product/marginal product (8) Total product refers to total output. Average product refers to output per unit of input. Marginal product refers to the change in output that results from a one-unit change in input.

total utility (6) See *utility.*

trade-off (26) Having more of one good at the expense of having less of another.

transfer earnings (14) A payment to an input which is equal to the amount necessary to elicit supply.

transfer payments (18,23) Government spending which is not made in exchange for a good or service. Transfer payments involve a transfer of income from the government to an individual or firm. Examples of such payments include unemployment compensation, welfare payments, and subsidies.

unemployment rate (22) The percentage of the labor force that is actively seeking employment but cannot find work.

uneven development/dual economies (36) In uneven development, a country has

both relatively well-developed and relatively undeveloped areas. In its extreme form, uneven development produces a dual economy: A modern, developed, wealthy sector, usually based on either tourism or a commodity export industry, is existing with a poor agrarian economy.

union (15) See *labor union.*

union shop (15) See *labor union.*

utility The satisfaction derived from the consumption of a good.

value of marginal product (14) See *marginal revenue product.*

variable costs Costs that vary with the level of output, such as payments for inputs. Average variable cost refers to variable cost per unit of output.

velocity of money (29) The rate of turnover of money. It equals GNP in current dollars divided by the money supply in current dollars.

vertical merger (10) See *merger.*

wager (23) See *factor income.*

windfall gains (14) Increases in economic rent which occur with no added effort or contribution by the owners of the inputs.

working class (37) A class of manual workers who produce, transport, sell, service, and maintain the material base of society. All other classes and groups depend on the working class for their needs.

x-inefficiency (10) Internal slack may result in the firm's costs for given levels of output rising above the lowest possible levels.

zero population growth (36) Occurs when the difference between the birth rate and the death rate is zero.

Index

Absolute advantage, 714
Accounting:
  choices for regulated utilities, 281
  international (*see* Balance of payments accounts)
  national, 483–503
  simple, of enterprises, 138–41
Accounting cost, defined, 138, 150 (*see also* Direct costs)
Accounts receivable, 140
Accumulation of capital (*see* Capital accumulation)
Ackley, Gardner, 626, 627 *fig.*
Addyston Pipe and Steel Company case, 254, 265 *tab.*, 267
Advertising:
  annual expenditures, 58
  as entry barrier, 223
Afghanistan crisis, 444, 474, 726
Africa, 688, 689, 733, 766, 777, 788
After-tax profit, 139
Age:
  demographic trends, 698
  and income, 404
Aggregate concentration, 241–42
Aggregate production function, 699, 700 *fig.*
Agricultural economics, 457–62
Agricultural products:
  inelasticity of demand and supply, 87–89
  price supports, 96–97
  selected, estimated elasticities of supply, 101 *tab.*
Agricultural sector:
  basic conditions, 457
  in China, 818, 819–20, 821, 822, 823
  energy scarcity, 458–59
  farm policies, 459–62
  long-term employment trends, 685–86
  mechanization, 693
  productivity, 700–702
  productivity gains, 350 *tab.*, 351, 457–58
  in Soviet Union, 809, 810–11
  in Third World, 773–74, 779–80, 784–85
  topsoil erosion and water shortage, 459
  and transportation industry, 705–6
Aid to Dependent Children (ADC), 404, 415, 659
Airlines industry:
  deregulation, 283
  development, 710–11
  public enterprises, 284 *fig.*
  response to fuel price increases, 173
Alabama, 21, 693, 694
Alaska, 465, 718, 755
Alcoa case, 255, 262 *tab.*
Alfa Romeo, 782
Allocative efficiency, 195–96 (*see also* Efficient allocation)
Allocative inefficiency, 209–11, 212
Altruism, utility analysis, 121
Aluminum Company of America, 227
American Broadcasting Company (ABC), 255–56

American Can Company, 255
American Economic Review, 12 *fn.*
American Electric Power Company, 268
American Revolution, 688, 694, 706, 780
American Stock Exchange, 344
American Telephone and Telegraph Company (AT&T), 6, 126, 147
  antitrust case, 252, 256, 262 *tab.*
American Tobacco Company, 227, 254, 262 *tab.*
Amtrak, 286
Anna Karenina (Tolstoy), 767
Antitrust and Trade Regulations Reporter, 261 *fn.*, 267
Antitrust policies, 249–70
  agencies and laws, 253–54
  economic criteria, 256–58
  economic effects, 258–59
  exempt organizations, 259 *tab.*
  toward existing concentrations, 257–58, 259–64
  history, 254–56
  toward mergers, 135, 257 *tab.*, 258, 264–67
  and minimum optimum scale, 171
  origins, 250–52
  toward price discrimination, 258, 268–69
  toward price fixing, 257 *tab.*, 258, 267–68
  standards, 252
Appropriate technology, 776–77, 784–85
Aquinas, Thomas, 8
Arbitrage, in currency market, 739–40
Aristotle, 7
Army Corps of Engineers, 383
Arrow, Kenneth J., 42
Artificial scarcity, of labor, 319, 324
Asia, 689, 766
"As if" proposition, 22, 107
Assets:
  on balance sheet, 139–40
  claims against, 140 (*see also* Liabilities)
  of commercial banks, 571, 582–83
  of Federal Reserve, 583–84
  financial, 570–71
  of firms, 53, 128–29 *tab.*
  household preferences, 19–20, 51
  of major sectors, 52 *tab.*
  and regulated price levels, 274, 277, 281
  "writing off," 142
Asset values, 343–49
Australia, 553, 774
Austria, 11, 284 *fig.*
Autolite, 258
Automation, disadvantages, 796–97 (*see also* Technological progress)
Automobile industry:
  autonomous innovation, 353
  emission standards, 432–33
  import quotas, 724–26
  market share concentration, 229
  market type, 200
  partial competition, 223 *tab.*
  public enterprises, 284 *fig.*
  ripple effects, 365
Autonomous inventions, 353
Autonomous shift, 524, 525–28, 530, 531–34, 535 (*see also* Multiplier)
Average cost, 194
Average fixed cost (AFC), 159 *tab.*, 160
Average product (AP), 154–58, 159 *tab.*
  and average variable cost, 160–61
Average propensity to consume (APC), 508–9
Average propensity to save (APS), 508–9
Average total cost (ATC):
  calculation, 159 *tab.*, 160, 162
  and economies and diseconomies of scale, 167–71
  and long-run equilibrium, 192–93
  in the long vs. short run, 164–67
  and marginal cost, 162–63
  and profit maximization, 180, 188
Average variable cost (AVC):
  and average product, 160–62
  calculation, 159 *tab.*, 160
  and marginal cost, 162–63
  and price, 186–87
  and profit maximization, 180
  and short-run equilibrium, 191
  and short-run supply curve, 184, 187–88
Babylonia, 7
Balanced budget multiplier, 644–45
Balanced growth, 706
Balanced reciprocity, 737–38
Balance of payments accounts, 740–53
  capital account, 741, 747–52
  components, 740–42
  current account, 740–41, 743–47
  equilibrium condition, 742–43
  exchange rate system, 753–61
  official settlements, 741–42, 757
Balance sheet, of enterprises, 138, 139–41
Bangladesh, 765–66
Bank deposit multiplier, 588, 591, 594
Bank deposits:
  as bank debt, 569, 571, 583
  in commercial banks, 571, 578, 583
  expansion and contraction, 584–94
  and money supply, 568–69
  in savings or thrift institutions, 572
  see also Bank reserves
Banking industry:
  institutions, 571–72, 578–84
  and money supply, 584–95
  recent innovations, 577–78
  see also Commercial banks; Federal Reserve
Bank reserves:
  components, 579–60 (*see also* Reserve deposits; Vault cash)
  and monetary base, 592
  regulations, 580–82 (*see also* Reserve requirements)
Barron's, 132
Belgium, 284 *fig.*
Benefit-cost analysis (*see* Cost-benefit analysis)
Bentham, Jeremy, 106
Berle, A. A., 127
Bias:
  in econometric diagrams, 50–51
  in present value analysis of natural resource use, 457
Bilateral monopoly, 325–26
Birth rate:
  and future world resources, 465
  in Third World, 777–78

Black markets, 96
Blacks:
  emancipation from slavery, 696
  exodus from South, 693
  housing discrimination, 408–9
  immigration, 688
  job discrimination, 406
  poverty among, 404
  unemployment rate, 6, 472, 479
Bloomsbury Group, 10
Blue-collar jobs, pay rates, 315–16
Bolshevik Party, 809, 817, 818
Bonds:
  asset values, 344–46, 347
  and cost of capital, 337
  interest rates, 342
  published price listings, 131, 132 *tab.*
  ratings, 341
  tax exempt, 397
  yields, 600, 601
Boston, 403, 479, 694
Bourgeoisie, 798
Brazil, 774
Bretton Woods Agreement, 753, 755–60, 761
Bribes, response to price ceilings, 96
Britain (*see* Great Britain)
British East India Company, 244
Broadbent, Sir William, 752
Brokers, 572–73
Brown Shoe—Kinney Shoe merger, 265 *tab.*
Bubble concept, 434, 435
Budget policy (*see* Federal budget; Fiscal policy)
Building industry (*see* Construction industry)
Built-in stabilizers, 517, 645–46, 666
Burger, Warren E., 256, 257
Burma, 788
Business cycle:
  changes in output, 35
  defined, 470
  early theories, 11
  long-term trends, 470–77
  Marxist analysis, 800–801
  and potential output, 36
  recent fluctuations, 477–81, 672–73
  stabilization policy debate, 36
  see also Economic instability
Business firms (*see* Firms)
Business investment (*see* Investment)
Business saving (*see* Gross business saving)
Business sector:
  in circular flow of goods and services, 101–2 (*see also* Circular flow)
  deficits, 497, 513, 515–16
  and household consumption, 510–11
  income and spending, 513–16
  percent of GNP received by, 496
  in Soviet economy, 812–13
  see also Industrial sector; Manufacturing sector
Business Week, 132, 256

California, 273 *tab.*, 425, 459, 692, 701, 733–34, 755, 774
Campbell Soup Company, 224
Canada, 4, 752
  agricultural specialization, 774
  tax revenues/GNP, 393 *tab.*
  trade patterns, 553, 720
  U.S. trade with, 730, 732 *tab.*, 733
Canal system, 707–8, 709 *fig.*
Capital (Marx), 8, 9, 12, 803
Capital, 331–54
  asset valuation, 343–49
  characteristics, 332
  costs of, 139, 335–36
  human, 316–18, 691
  as input to production, 20–21, 137, 153
  investment decisions, 142, 333–37
  management of, 136, 137
  market demand, 337
  market supply, 337–38
  price of, and least-cost production, 171–72
  and productive capacity, 5
  productivity of, 332
  real or physical vs. money or portfolio types, 331–32
  return on, 141, 332, 338–43
  and technological change, 349–53
Capital account, 741, 747–52
Capital accumulation:
  and economic growth, 699–711
  in manufacturing, 704–6
  Marxist view, 797–803
  and population growth, 778
  in Third World, 775–77, 782–83
  in transportation industry, 706–11
Capital goods, 331–33
  vs. intermediate goods, 486
Capital intensity:
  defined, 148 *fn.*
  and elasticity of supply, 77–78
  and industrial growth, 705
Capitalism:
  Chinese reforms distinguished from, 824
  in classical economics, 8
  defined, 21
  Marxist view, 13, 791–804
  vs. socialism, in the Third World, 781–82
  stability issues, 476
Capitalist class, 798
Capitalist mode of production, 793
Capital services, 485
Capital stock, 30
Caribbean region, 766
Carlson, Chester, 405
Carnegie, Andrew, 405
Carnegie-Mellon University, 405
Cartels:
  defined, 234
  in nuclear fuels, 463–64
  in petroleum (*see* Organization of Petroleum-Exporting Countries (OPEC))
Carter administration, 676–81
Cash, on balance sheet, 140 (*see also* Coins; Currency; Vault cash)
Cash crop agriculture, 779, 785
Central America, 444, 733, 766
Central planning, in Soviet Union, 808–17
Certificates of deposit, 568–69
Chad, 772
Charity:
  provision of public goods, 385
  utility analysis of contributions, 121
Charts (*see* Diagrams)
Check clearing, 579–80, 582

Checking deposits:
in commercial banks, 571, 578
money supply component, 568
reserve requirements, 581
technical name, 569
see also Bank deposits
Chemical Bank, 588
Chesapeake & Ohio Canal, 707
"Chicago" (Sandburg), 731
Chicago, 478, 694, 695
Chicago School, 13
Chicanos, 479
Chile, 781
China, 7, 658, 817–25
agricultural communes, 784
capital accumulation, 783
Cultural Revolution, 823–24
disguised unemployment, 775
emergence as socialist, 803
future of, 824–25
Great Leap Forward, 821–23
investment/output ratio, 776
Marxian economics, 10, 792
population problems, 777, 817–18
reconstruction and first five-year plan, 819–21
Revolution of 1911, 818
and Soviet Union, 788
and Third World, 766, 787
Choices (*see* Economic choices)
Chrysler Corporation, 147, 229, 258, 517, 649, 725, 726, 730
Churchill, Winston, 824
Circular flow:
autonomous shifts and induced changes, 524 (*see also* Multiplier)
demand and supply, 18–19, 23–24, 33–34
equilibrium of, 505–520
financing, 565–75
measurement, 488–95
two-sector and four-sector, 485–88
Civil Aeronautics Boards (CAB), 273 *tab.*, 274, 283
Civil War:
and business cycle, 470, 472
and economic growth, 250, 689, 692, 693
and education trends, 698
and gold standard, 755
and railroad system, 709
Vietnam War compared to, 669
Clark, Colin, 393
Classical economics, 8, 12
Classical liberals, 13
Class structure:
in China, 823–24
Marxist view, 797–99
in Soviet Union, 817
Clayton Act (1914), 254 *tab.*, 255, 257 *tab.*, 264
Cleveland, 478, 600, 695
Clinical analysis, 14
Clorox Chemical Company, 266
Closed shop, 320, 321
Coal industry, public enterprises, 284 *fig.*
Coins, 569–70
Cold War, 475, 729
Collective bargaining (*see* Labor unions)
Collectivization:
in China, 819–21
in Soviet Union, 810–11, 812, 815, 819
Colleges, public financing issues, 422–23, 425–28
Collusion:
antitrust cases, 267–68
antitrust policy, 254 *tab.*
vs. competition, 232–33
conditions favorable toward, 233–34
types of, 234–35
Colonialism, 780–81
Colorado, 273 *tab.*
Commercial banks:
assets and liabilities, 569, 582–83
chartering and regulation, 578–82 (*see also* Reserve requirements)
creation of bank money, 584–91
five largest, 129 *tab.*
role in banking system, 571, 578
see also Bank deposits; Banking industry; Bank reserves
Commodities:
foreign trade, 731–33
heavily taxed, 94
market prices, 56 *tab.*
Marxist view, 793–95
published price listings, 131, 133 *tab.*
Common-property resources, 369, 456
Common stock (*see* Stock)
Communes, 821–22, 823
Communications industry, productivity growth, 350 *tab.*, 351
Communist Manifesto (Marx and Engels), 9, 794
Communist Party:
Chinese, 818–19, 820
Soviet, 809, 811, 813
Companies (*see* Enterprises; Firms)
Comparative advantage:
as acquired characteristic, 785
of "cheap foreign labor," 725
and economies of scale, 718
and job selection, 307
law of, 714–15
reflected in U.S. trade, 733
specialization and gains from trade, 715–17
of Third World nations, 781, 785, 786
and transportation costs, 717–18
Comparative economic systems, 808
Comparative static analysis of prices, 57
Competition:
defined, 180
degrees of 221–47 (*see also* Partial competition)
and efficient allocation, 194–96, 358–64
entry barriers (*see* Entry barriers)
ideal model (*see* Pure competition)
from imports, 244
limits on efficiency of, 196, 364–70
and maximization of consumer surplus, 210
vs. monopoly, 200
natural, 170
nature of, 180–81
of oligopoly firms, 232
perfect (*see* Perfect competition)
policies that reduce, 259 *tab.*
policies to increase (*see* Antitrust policies)
public school voucher proposal, 424–25
recent market share trends, 243–44

for regulated utilities, 282–83
and rivalry, 181
Schumpeter's theory, 231
unfair, 204
U.S. faith in, 249
Complementary goods:
cross-elasticity of demand, 70–71
defined, 58
and derived demand, 120
Comptroller of the Currency, 578
Computer industry, 261–63, 269
Concentration:
aggregate, 241–42
and antitrust policy, 259–67
and economies of scale, 239–40
in monopolistic competition, 240
and oligopolistic collusion, 232, 234
and oligopoly characteristics, 229–30, 231
recent trends, 242–44
Concessional loans, 787
Conglomerate firms:
characteristics, 126–27
effect on competition, 244–46
Conglomerate mergers:
antitrust policy, 256, 257 *tab.*, 258, 265 *tab.*, 266–67
recent trends, 130–35, 245
Conservation, 448–57
Conservative economists, 13, 36
Construction industry:
and 1970s business cycle, 478
productivity growth, 350 *tab.*, 351
sensitivity to interest rates, 609
Consumer demand:
basis for, 20
and derived demand, 297
and interest rates, 608
see also Demand; Individual demand; Planned demand
Consumer durable goods, 333
Consumer goods, markets for, 23
Consumer price index (CPI), 499–501, 502, 558, 675
Consumer prices:
fluctuations in 1970s, 479
long-term trends, 472–73
see also Prices
Consumer Product Safety Commission (CPSC), 435 *tab.*, 436–37
Consumers:
rational choices, 105–7, 120–21
share of sales taxes, 92–94
see also Household sector
Consumer surplus:
and marginal utility, 117–18
and market demand, 120
Marshall's contribution, 106
reduction in, due to monopoly, 210
Consumption (C):
in calculation of sectoral surpluses or deficits, 497, 498
and disposable income, graphing, 41–43, 46–48
and equilibrium theory, 506–13
and final demand, 494, 495
household patterns, 19–20, 104–5
propensity for, 506–10
of social vs. private goods, 375–76
Consumption function, 508, 509 *fig.*
and multiplier effect, 525–26
Contracts:
exclusive and tying, antitrust policy, 254 *tab.*
government, 204
Conventional wisdom, 792
Convertible currencies, 756–57
Coolidge, Calvin, 484
Cooperatives, 135
in China, 821–22
Corporate income tax, 396
Corporate profits:
of largest firms, 128–29 *tab.*
long-term trends, 342–43
see also Profits
Corporate securities, ownership, 402 (*see also* Bonds; Stock)
Corporations, 125–27 (*see also* Firms)
Correspondent banks, 579, 580
Cost-benefit analysis:
antitrust standard, 252
of capital goods, 332–37
and choices of enterprises, 136–37
and economic approach, 14
of environmental protection, 430–32, 434, 435 *fig.*
of finding and harvesting natural resources, 453–54
for government choices, 32
of marginal input, 290
and marginal utility, 116, 117
and opportunity cost, 25 (*see also* Opportunity cost)
of public expenditures, 378–83
of worker and consumer safety programs, 436–37
see also Efficient allocation; Present value analysis
Cost-of-living adjustment (COLA), 500, 558
Cost-plus pricing, 439–40
Cost-push inflation:
defined, 540
vs. demand-pull inflation, 551
and labor costs, 544
Costs, 147–74
accounting, 150
accounting vs. economic, 138
average, 194
average fixed, 159 *tab.*, 160
average total, 159 *tab.*, 160, 162–63, 164–71, 180, 188, 192–93
average variable, 159 *tab.*, 160–63, 180, 184, 186–88, 191
of capital, 139, 335–38
defined, 149
direct, 150, 151–53, 297–98, 453–54
discounted (*see* Present value analysis)
and economic profits, 151–53
effects of regulation on, 281–82
explicit, 150
external vs. social, 364–66
fixed vs. variable, 159
of foregone alternatives, 149–50 (*see also* Opportunity cost)
of funds, 599, 634
implicit, 150
imputed, 150
of inputs, 540–49
and productivity, 153–73, 549–51
sunk, 150, 151
and technology, 148–49
and transfer earnings, 299–300
Cotton Belt, 694
Council of Economic Advisers (CEA)
conduct of fiscal policy, 625, 626, 627–29, 630

and high-employment budget, 646–47
historical attempts at stabilization, 665, 666, 667, 673, 675, 683
Council on Environmental Quality, 429
Craft unions, 322, 323–24
Cream skimming, 282
Creative destruction, theory of, 231
Credit, cost of, 599
Credit crunch of 1966, 669–71, 673
Credit unions, 572
Cross-elasticity of demand, 70–71
Crowding out phenomenon, 616–17, 636
Reagan program to counteract, 682
Cuba, 408
Cultural Revolution, 823–24
Cultural values, and efficient allocation, 368
Culture, effect of monopoly on, 213
Currency:
convertible, 756–57
devaluations, 759
as drain on deposit creation, 590–91
as Federal Reserve debt, 569, 584
international flows (*see* Balance of payments accounts)
and monetary base, 592
money supply component, 568
Currency markets, 738–40
exchange rates, 730, 731, 753–61
official settlements, 742, 757
Current account, 740–41, 743–47
Current assets, 140
Cyclical unemployment, 552–54

Dark Ages, 8
Das Kapital (see *Capital*)
Dealers, 572–73
Death rate, in Third World, 777–78
Debt:
cost of financing, 337, 338
money as, 569–70
negotiable, 570
Deception, in econometric diagrams, 50–51
Decision making:
in households, 18, 19–20
individual, 117
and innovation, 352–53
and investment, 333–34
see also Cost-benefit analysis; Economic choices
Defense spending (*see* Military spending)
Deficits:
in business sector, 497, 513, 515–16
in government sector, 475, 497, 517, 649, 683
see also Surpluses and deficits
Deflation, 501–2, 620
Demand, 58–71
for capital, 337
cross-elasticity of, 70–71
derived, 120, 297, 311
diminishing marginal effect, 27
elastic, 86–87
elasticity and total revenue, 66–67
and income, 34–35
income elasticity, 67–70
individual vs. total, 104, 118–20
inelastic, 87–90
influences on, 58–59
for inputs, 290–97
interaction with supply, 78–80
for labor, 311–13, 324
for money, 599–604
price elasticity, 63–67
vs. quantity demanded, 61–63
for social goods, 376–77
see also Demand curve; Demand and supply; Final demand; Individual demand; Market demand; Planned demand
Demand and supply, 55–83, 85–102
for agricultural products, 460
for college education, 425–27
for currencies, 739
effects of elasticities on market outcomes, 86–91
household-enterprise interaction, 18–19
and inflation, 540
interaction of, 78–80
interference with market process, 91–98
for investment funds, 34–35
and market price, 23, 55–57
measuring, 98–101
for money, 604–6
for soldiers, 443–44
tendency toward equilibrium, 32
see also Demand; Supply
Demand curve, 59–61
of dominant firm, 222–23
for free vs. scarce goods, 112–13
graphing conventions, 41–43
of individuals 104, 107, 110–11, 116–17
intersection with supply curve, 78–80
kinked, of oligopoly, 236–39
in monopolistic competition, 240–41
in monopoly, 200–202
in perfect competition, 181–82
in pure competition, 180
see also Elasticity of demand
Demand deposits (*see* Checking deposits)
Demand management:
as Keynesian emphasis, 681
with supply side problems, 676
Demand-pull inflation, 540, 552–57
plus expectational inflation, 562
due to slowdown in productivity, 551
Democracy, effect of monopoly on, 213
Demographic changes, 695–99 (*see also* Population growth)
Demographic transition, 777–78
Denver, 479, 695
Dependency ratio, 691, 698
Dependent variable, 41, 46–47
Deposit contraction, 588–90
due to high discount rate, 594
due to open market operations, 593
Deposit creation:
complications, 589, 590–91
due to lowered reserve requirements, 594
multiplier effect, 584–88
Depository institutions:
and creation of bank money, 584–91
Federal Reserve loans to, 584
regulation and control, 581–82

see also entries under Bank
Depository Institutions Deregulation and Control Act (1980), 579, 580, 581–82
Deposits (*see* Bank deposits; Reserve deposits)
Depreciation:
effect on asset values, 140
and gross business saving, 496, 513–14
on income statement, 139
in input-output accounting, 494–95
Depressions:
hallmark of, 471
in Keynesian theory, 11
see also Great Depression
Deregulation, 282–83
Derived demand:
defined, 120
for labor, 311
method of computing, 297
Detroit, 478, 694, 695
Developing nations (*see* Third World economies)
Development strategies:
of China, 819–23
of Soviet Union, 809–11
of Third World, 781–88
Diagrams, 40–50
of distributions, 49–50
economic models, 44–48
functions, 40–41
of linear equations, 41–44
problems of bias and deception, 50–51
three main types, 41
of time series, 48–49
Dialectic of history, 803
Diminishing marginal effect, 26–27
of market disequilibrium, 32
and production-possibility boundary, 29, 30 *fig.*
and public choice, 32–33
Diminishing marginal returns (*see* Law of diminishing marginal returns)
Diminishing marginal utility (*see* Marginal utility)
Direct costs:
defined, 150
and economic profit, 151–53
and input market supply curve, 297–98
of natural resources, 453–54
vs. opportunity costs, 149–50, 151
Direct investment:
international, 748, 750 *fig.*, 751–52
in Third World, 787
Discounted future values (*see* Present value analysis)
Discount rate, 594, 654
Discrimination
in employment, 406–8
in housing, 408–9
programs to combat, 413–16
Diseconomies of scale, 167–71
Disguised unemployment, 775
Disposable income (YD):
in calculation of sectoral surpluses or deficits, 497
and consumption, graphing, 41–43, 46–48
and distribution of GNY, 496
and equilibrium GNP, 510–11, 512
and gross business saving, 513, 516
and propensity to save or consume, 506–10
taxation effects, 410–11
Distribution, economic choices, 2–3 (*see also* Income distribution)
Distributions, in econometric diagrams, 41, 49–50
Distribution sector, 350 *tab.*
Disutility, 108, 306
Diversification:
Diversification:
of large corporations, 126–27
and risk reduction, 341
Dividends:
and asset values, 347–48
and cost of capital, 337
and income statement, 139
long-term trends, 342
vs. retained earnings, 140–41
Dividend yield, 348
Dollar:
exchange values, 762 *fig.*
international market, 740–43
as international reserve currency, 757–60
Dominant firms, 222–29
antitrust policy, 258, 259–64
causes, 224–29
characteristics, 222–23
instances and effects, 223–24, 225 *tab.*, 228 *tab.*
and market type, 201 *tab.*
predatory pricing, 269
price discrimination, 216
recent market share trends, 243–44
Dos Passos, John, 471
Dow Jones Industrial Average, 343, 345 *fig.*
Dual economy, 774, 780
Duke, James B., 405
Duke University, 405
Duopoly, 236
DuPont Corporation, 245, 262 *tab.*, 265 *tab.*, 266
DuPont family, 403, 405

Eastman Kodak Company, 171, 229, 263
East (U.S. region), 690, 709–10
Econometrics, 39–54
comparative static analysis, 57
distributions, 49–50
fitting techniques, 47
linear equations, 41–44
model building, 44–48
Nobel Prize winners, 42–43
problems of bias and deception, 50–51
stock vs. flows, 51–53
time series, 48–49
use of diagrams, 40–41
Economic analysis, 7–14
in ancient times, 7–8
classical, 8, 12
comparative economic systems, 808
of demand and supply, 85–102 (*see also* Demand and supply)
economic approach, 13–14
and economic choices, 2
focus on markets, 23
literature, 12–13
methods and measurements (*see* Econometrics)
neoclassical, 8–12
positive and normative, 4
of utility and demand, 104–20
see also Economics; Economists
Economic choices:
in classical economics, 8
and efficient allocation, 359

of enterprises, 20, 136–37, 138
as focus of economics, 1–2
and general equilibrium, 356–58
of households, 18, 19–20
individual decision making, 117
intergenerational, 454–55
of investors, and equalization of return, 348
in job selection, 307
marginal conditions, 25–26
maximizing behavior, 22
of monopolist, 204–9
opportunity cost, 25 (*see also* Opportunity cost)
and production-possibility boundary, 27–31
production questions, 2–3, 19, 24–25
of production technologies, 148–49
public, 32–33
and public school monopoly, 423–25
rational, 105–7, 120–21
restriction of, due to monopoly, 212–13
in the short and long run, 166–67
taxation effects, 387–89
Economic cost, 149, 150 (*see also* Opportunity cost)
Economic development (*see* Development strategies; Third World economies)
Economic goals, 3–4
Economic growth, 301–28
in agriculture, 700–702
historical significance, 686
and investment levels, 30
long-term vs. short-term, 685–86
in manufacturing, 702–6
population and demographic factors, 465–66, 688–99
in Soviet Union, 815
theory of, 699–700
in transportation industry, 706–11
U.S. trends, 470
Economic instability, 476, 616
Economic models, 41, 44–49
Economic obsolescence, 352–53
Economic planning, in Soviet economy, 808–17
Economic profit, 151–53, 338
Economic Recovery Tax Act (1981), 682
Economic rent:
and elasticity of supply, 299–301
Marshall's contribution, 106
and natural resources, 448
and price of farmland, 460–61
taxing of, 301–2
Economic Report of the President (1982), 683
Economics:
agricultural, 457–62
applied fields, 222
defined, 1–2
as the "dismal science," 448, 688
of energy, 462–65
evolution of, 12
Nobel Prize, 42
political process, 373
principles, 17–37
see also Economic analysis; Economists; Macroeconomics; Microeconomics
Economic sectors, 4–5
in the Soviet Union, 811–13
stocks and flows, 51–53
Economic Stabilization Act, 674, 675
Economic systems, 2–7
circular flow of goods and services (*see* Circular flow)
distribution of income and wealth, 6–7
in economic analysis, 13–14
and economic goals, 3–4
influence of scarcity and choice, 2–3
major sectors (*see* Economic sectors)
positive and normative analysis, 4
productivity factors, 5–6
subject of macroeconomics, 470
tendency toward equilibrium, 32
Economies of scale, 167–71
in agriculture, 701, 784
antitrust criteria, 261
as cause of dominance, 224–27
and comparative advantage, 718
as entry barrier, 223
and monopoly power, 203, 218, 368
of oligopoly, 239–40
and regulation of utilities, 272, 273, 275
Economists:
debate over government stabilization policy, 36, 476
desirable attributes, 14, 40
generalists vs. specialists, 11–12
major figures, 8–11
Nobel Prize winners, 42–43, 357, 598, 687
publications, 12–13
schools and groups, 13
Edison, Thomas, 217
Education:
demographic trends, 698–99
expenditures on, 422–23
private benefits, 316–19, 420–21
public benefits, 421
public colleges, 425–28
public school monopoly, 423–25
in Third World, 778–79
Efficiency:
antitrust standards, 252
conditions for, 194–96
economies and diseconomies of scale, 167–71 (*see also* Economies of scale)
and financing of public colleges, 425–28
vs. growth, in Soviet economy, 815
of income distribution, 405
and monopoly power, 203
and production-possibility boundary, 27–31
in public action, 32, 374
of public enterprises, 286–87
and regulated price levels, 276–77, 281
X-level, 165 (*see also* X-efficiency)
Efficient allocation:
and college financing, 427
conditions of, 195–96, 358–61
effect of farm price supports, 460
and general equilibrium, 355
limits, 196, 364–70
of natural resources, 448–49
see also Efficiency

Effort, and income distribution, 405
Effluent fees, 434
Egypt, 7
Eisenhower, Dwight D., 438
Elastic demand, defined, 65
Elasticity:
  defined, 63
  main types and ranges, 82 *tab.*
  and market outcomes, 86–91
Elasticity of demand:
  cross-elasticity, 70–71
  effect on future world resources, 466
  for an input, 294–95
  and incidence of taxation, 387
  with kinked demand curve, 236–39
  for labor, 314, 320–21
  and monopoly conditions, 208, 211
  price discrimination based on, 213–16
  and regulation, 273
  relative to income, 67–80, 82 *tab.*
  relative to price, 63–67, 82 *tab.*
  and sales tax burden, 92–93
  trade effects, 745, 746–47
  see also Demand curve
Elasticity of supply, 75–78
  and economic rent, 299–301
  effect on future world resources, 466
  and incidence of taxation, 387
  of individual laborers, 308
  for labor market, 314–15
  and sales tax burden, 93–94
  see also Supply curve
Elastic supply, defined, 75
Elderly, poverty among, 404
Electrical Equipment Conspiracy, 265 *tab.*, 267–68
Electric power companies:
  as monopolies, 217–18
  productivity growth, 350 *tab.*, 351
  public enterprises, 284 *fig.*
  regulated price levels, 272, 277, 278–81
  see also Utilities
Elements of Pure Economics (Walras), 357
Ellis Island, 690
Employment:
  in agriculture, 700
  discrimination, 406–8
  effects of labor unions on, 322–23
  equal opportunity programs, 413–16
  growth during 1970s, 677–78
  in manufacturing, 703
  and minimum wage law, 416–17
  see also Full employment; Labor; Unemployment; Work
Employment Act (1946), 624–25, 626
Energy, economics of, 462–65 (*see also* Fuels; Oil industry; Oil prices)
Engels, Friedrich, 797
England (*see* Great Britain)
Enterprises, 123–45
  accounting concepts, 138–41
  choices and outcomes, 136–37
  defined, 135–36
  diversity in size, 6
  functions in economic system, 18–19, 20
  income requirements, 20
  input categories, 20–21
  inputs, outputs, and production, 137–38
  maximizing behavior, 21–22
  new, a case history, 142–45
  patterns, 124–35
  public, 20, 135, 283–87
  success indicators, 141–42
  see also Firms; Private enterprise
Entrepreneurship, 342, 352
Entry barriers:
  absence of, as condition of competition, 181, 240, 241
  in craft unions and professional groups, 324
  foreign trade protectionism, 720–27
  due to monopoly, 202, 212–13
  of oligopoly, 232
  types and causes, 223
Environmental protection, 428–35
  automobile emission case, 432–33
  cost-benefit analysis, 430–32
  issues and programs, 428–30
  steel industry case, 433
  use of incentives, 433–35
Environmental Protection Agency (EPA), 429, 432, 433, 434
Equal advantage, principle of, 600–601
Equal Employment Opportunity Commission, (EEOC), 413–16
Equalization of returns, 348
Equal Opportunity Act (1964), 413
Equilibrium:
  of marginal utilities and prices, 113–17
  microeconomic principle, 31–32
  of prices and quantity, 78–80
  short- and long-run, under perfect competition, 190–93
  see also General equilibrium
Equilibrium theory:
  analysis of circular flow, 505–20
  multiplier effects, 523–37
Equities, as financial assets, 570
Equity (*see* Fairness)
Equity capital:
  on balance sheet, 140
  cost of, 337, 338
  and profitability, 141
Erie Canal, 707
Essay on the Principle of Population (Malthus), 8, 688
Estate taxes, 396
Ethics, and use of natural resources, 454–55 (*see also* Values)
Ethiopia, 772
Eurodollars, 760
Europe, 8, 11, 688–92, 729, 751 (*see also* Western Europe)
Excess profits:
  of dominant firms, 224
  of oligopoly firms, 232–33
  regulation to prevent, 272
Excess reserves:
  and creation of bank money, 584–88, 589–91, 592
  and interest rates, 606–7
  due to lowered reserve ratio, 594
Exchange:
  market system, 22–23
  money as medium of, 567
Exchange rates:
Bretton Woods system, 730, 755–60
  and currency market, 738–40

floating rate system, 731, 753, 760–61
gold standard, 753–55
Excise taxes, 413 (*see also* Sales taxes)
Expectational inflation, 558–62, 631–32
Expectations:
and asset values, 343, 346–47
effect on demand, 58
effect on prices, 455, 462 (*see also* Expectational inflation)
effect on supply, 72
of persistent inflation, 631–32
rational, 631, 683
Explicit costs, defined, 150 (*see also* Direct costs)
Exploitation of labor, 796
Export promotion, 786
Exports:
net, 497, 518
restrictions on, 726–27
since World War II, 728–31
see also Foreign sector; International trade
Export surplus, 730, 731
External effects:
as limits to efficient allocation, 364–69
and public policy, 32–33, 375 *tab.*, 377–78
and rate of use of natural resources, 369, 457
External funds, 337
Exxon Corporation, 126, 141, 203, 217

Factor income, 485
Factor services, 485, 487
Factors of production, 20–21 (*see also* Inputs)
Fairness:
and efficient allocation, 367–68
equity in public sector, 374
and financing of public colleges, 425–28
Family planning, in Third World, 783–84
Family size, and income, 404–5
Farmers, poverty among, 404
Farming (*see* Agricultural sector)
Farmland:
effect of price supports on value, 460, 462
erosion of topsoil, 459
land bank program, 461
percent of total land, 700
Farm policy, 459–62
Fascism, 213
Fast-food restaurants, case study of partial competition, 223 *tab.*
Federal budget:
built-in stabilizers, 645–46
and business investment, 665–67
components, 642–43
high-employment, 646–49
legislative enactment, 525–26
multiplier effect of changes in, 643–45
and national debt, 649–52
See also Fiscal policy
Federal Communications Commission (FCC), 273 *tab.*, 274
Federal Deposit Insurance Corporation (FDIC), 579
Federal Energy Regulatory Commission, 273 *tab.*, 274
Federal Farm Loan program, 396
Federal funds:
defined, 586
and discount rate, 654
and money supply, 590
Federal government, long-term growth trends, 474–77 (*see also* Government; United States)
Federal Reserve (Fed):
assets and liabilities, 583–84
Board of Governors, 580, 582, 612
central bank ownership, 584
conduct of monetary policy, 612, 652–56
currency as debt of, 569, 570
discount rate, 594, 654
and equilibrium interest rate, 605, 606–8, 612–13
and floating exchange rate, 760
and gold standard, 754
historical attempts at stabilization, 669–70, 672, 679–80, 682
history and functions, 578
independence, 624, 650
influence on multiplier, 645
member and nonmember banks, 578–79, 580–81
open-market operations, 591, 592–93, 594, 612–13, 653
as owner of national debt, 583, 650
recommended monetarist strategy, 637–38
regulation of banking industry, 578–82 (*see also* Reserve deposits; Reserve requirements)
regulation of money supply, 568, 578, 579, 581, 583–84, 589, 591–94, 599, 616, 636, 637–38
response to inflation, 618
support of government securities, 649
Federal Trade Commission Act, 254 *tab.*
Federal Trade Commission (FTC), 253, 255, 256, 257 *tab.*
Dupont case, 262 *tab.*
Proctor & Gamble case, 265 *tab.*, 266
Xerox case, 262 *tab.*, 263–64
Fiat Company, 782
Fiat money, 569
Fibers and textiles:
market prices, 56 *tab.*
published price listings, 133 *tab.*
Figures (*see* Diagrams)
Final demand:
components, 494–95
deflation, 501
vs. derived demand, 120
and GNY, 495
and value added, 490–91
value of goods delivered to, and GNP, 488–89, 492–94
Financial assets, 570–71
Financial institutions, 566
and production sectors, 573
productivity growth, 350 *tab.*, 351
public enterprises, 283
types, 570–73
Financial instruments, rate of return, 598–99
Financial intermediaries, 572
and flow of funds, 573–74
as owners of national debt, 650–51
Financial markets:
defined, 23
function, 566

supply and demand factors, 34
see also Stock market
Financial panic, threat in 1966, 670–71
Fines, 383
Finished goods, 137
Firms:
conditions of efficient allocation, 358 *tab.*
corporations, 125–27
derived demand, 120
dominant (*see* Dominant firms)
economic choices, 2
financial assets, 570–71
influences on supply, 72
investment demand, 34–35
news stories, 131
public (*see* Public enterprises)
scope of term, 123
in Soviet economy, 812–13
stocks and flows, 51–53
supply of capital to, 338
in two-sector economy, 485
see also Business sector; Enterprises
Fiscal policy:
of Carter administration, 678–79
conduct of, 624–26
coordination with monetary policy, 656–60
discretionary, 638
of Johnson administration, 669, 671–72
of Kennedy administration, 665–68
limitations as stabilizer, 638
of Nixon administration, 673, 676
principles, 642–49
see also Stabilization policy
Five-Year Plans:
in China, 819–21
in Soviet Union, 811, 814
Fixed assets, 140
Fixed cost:
average, 160
defined, 159
effect on short-run supply curve, 190
total, 160
Fixed inputs, 153
Florida, 457, 692
Flow of funds, 573–74, 634
Flows, vs. stocks, 51–53 (*see also* Circular flow)
Food and Drug Administration (FDA), 435, 436–37
Foods:
market prices, 56 *tab.*
published price listings, 133 *tab.*
see also Agricultural products
Forbes, 132, 256
Forced industrialization, 809–10
Forces of production, 798, 823
Ford administration, 675, 676, 678
Ford Motor Company, 51–53, 229, 258, 700, 730
Forecasting, and natural resource use, 456–57
Foregone alternatives, 149–50 (*see also* Opportunity cost)
Foreign aid:
and farm policy, 461
to Third World, 787–88
Foreign currencies:
exchange rates, 131, 134 *tab.*, 730, 731, 753–61
Federal Reserve holdings, 584
markets, 738–39
U.S. government transactions, 741–52
Foreign investment:
capital account, 747–52
in Third World, 786–88
Foreign sector:
in circular flow of goods and services, 486–87, 488
composition of trade, 731–34
and equilibrium GNP, 517–18
importance to U.S. economy, 727–31
influence on economic system, 18
in Soviet economy, 813
surpluses and deficits, 497, 518
see also International trade
Fortune, 132, 256
Forward pricing, 455
France, 284 *fig.*, 393 *tab.*, 729
Franchises:
monopolistic, 203–4
of regulated utilities, 272, 274
Franklin, Benjamin, 443
Free goods:
and consumer surplus, 118
and marginal utility, 112–13
Free-market economists, 13
Free trade:
barriers to, 720–27
Ricardian argument, 717
French physiocrats, 8, 302 *tab.*
French Socialists, 8
Frictional unemployment, 552
Friedman, Milton, 42, 424, 476, 481, 598, 637, 638, 656, 657, 676, 682, 761
Frisch, Ragnar, 42
Fuels:
conservation by airlines, 173
economies of use, 462–65
effect of scarcity on farming, 458–59
elasticity of demand, 466
price controls, 95
see also Oil industry
Full employment:
as government policy, 624
and money supply, 619–20
and national debt, 652
Full Employment Act (1945), 624
Funds:
cost of, 599
federal, 586, 590, 654
flow of, 573–74, 634
internal, 337
Future (*see* Expectations)
Future values, discounted (*see* Present value analysis)

Galbraith, John Kenneth, 260, 792
Gang of Four, 824
Gas industry, public enterprises, 284 *fig.*
Gasoline taxes, 387, 395
General Electric Company, 227, 245, 265 *tab.*, 268, 269
General equilibrium, 355–70
adjusting toward, 356–58
conditions, 358–61
Invisible Hand concept, 361, 364, 367, 370, 456
limits on competitive efficiency, 364–70
ripple effects, 356, 361–64, 365
General Foods Company, 245
General Motors Corporation, 2, 6, 22, 135, 171, 200, 203, 224, 229, 245, 259, 265 *tab.*, 266, 730, 808
General Theory of Employment, Interest, and Money (Keynes), 10, 11, 12, 681
Geography:
and income levels, 404

and market definition, 57, 261
and ripple effects, 364
and unemployment, 478–79
George, Henry, 301
Georgia, 273 *tab.*
Germany, 688, 689–90, 629, 809 (*see also* West Germany)
Gillette Company, 227
GNP (*see* Gross national product)
GNP deflator:
derivation, 501–2
during 1970s, 562
function, 499
GNP effect, 611–12
GNP gap, 627–30
GNP multiplier (*see* Multiplier)
Gold:
market value, 56 *tab.*, 57
published price listings, 133 *tab.*
Goldberg, Rube, 623, 624 *fig.*
Gold Rush, 692, 755
Gold standard, 753–55
Goods:
complementary, 58, 70–71
costs of producing, and supply, 71–72
with inelastic supply, 87
intermediate (*see* Intermediate goods)
market definition, 57, 261
necessities, 67
normal vs. inferior, 69–70
related, 72
scarce vs. free, 111–13
social vs. private, 375–76
substitutable (*see* Substitutable goods)
Goods and services:
circular flow (*see* Circular flow)
national input-output accounting, 488–95
as output of enterprise, 138
price indexes, 499–501
selected, price elasticities of demand, 100 *tab.*
see also Goods
Gosplan, 814
Government (*see* Local government; State governments; United States)
Government contracts, 204
Government debt, 649–52 (*see also* Government securities)
Government intervention, views of various schools of economists, 11, 13
Government policy:
conglomerate pressures on, 245–46
effect on competition, 244
farm policy, 459–62
and monopoly power, 203–4, 249–70 (*see also* Antitrust policies)
see also Public policy
Government purchases, 385–86
in calculation of sectoral surpluses or deficits, 497
and equilibrium GNP, 516, 517
and final demand, 494, 495
and stabilization policy, 643–45
trends, 392, 393 *fig.*, 394, 474–75
see also Government spending
Government regulation (*see* Regulation)
Government sector:
in circular flow of goods and services, 486–88
criteria for economic choices, 32–33
and equilibrium GNP, 516–17
expenditures (*see* Government spending)
financial assets, 570–71
influence on economic system, 18
interference with market process, 94–98
percent of GNP received by, 496
public goods provided by, 113
in Soviet economy, 813
stocks and flows, 52 *tab.*, 53
surpluses and deficits, 475–77, 497, 517, 683
in U.S. economy, 4
Government securities:
demand for, 649
Federal Reserve assets, 583–84, 653
open-market operations, 591, 592–93, 594, 612–13, 653
ownership, 592
and reserve requirements, 580–81
Government spending:
during Carter administration, 678–79
components, 385–86
inflationary vs. multiplier effects, 535–37
long-term trends, 475–77
Reagan reductions, 682
see also Government purchases; Military spending; Transfer payments
Grain Belt, 694, 705, 706
Grains and feeds:
labor productivity trends, 701
market prices, 56 *tab.*
published price listings, 133 *tab.*
Soviet wheat purchases, 553–54, 656
Gramlich, Edward M., 417 *fn.*
Graphs (*see* Diagrams)
Great Britain:
classical economics, 8
comparative advantage, 785
Corn Laws, 717, 724
currency devaluation, 759
development strategy, 782
free trade movement, 717
heyday of capitalism, 368
immigration from, 688, 692
neoclassical economics, 11
public enterprises, 284 *fig.*
trade effects of World War II, 729
trade with U.S., 744–47
Great Depression:
and antitrust policy, 255
bank failures, 578, 579
effect on immigration, 692
excess reserves, 589
expansion of federal government, 475
expansion of national debt, 649
and gold standard, 755
Keynes's analysis, 10, 11, 524
Marxist view, 476
post–World War II expectations, 624
scope of, 470–71
unemployment rate, 472, 479
Vietnam disruption compared to, 669, 673
Great Lakes region, 692, 695, 704, 707
Great Leap Forward, 821–23
Great Plains region, 6, 459
Great Proletarian Cultural Revolution, 823–24

Great Society, 392, 394
Greece, 7
Greenfield Village Museum, 700
Green revolution, 785
Gross business saving (SB):
and business deficit, 515–16
in calculation of sectoral surpluses and deficits, 497
components and uses, 513–14
and distribution of GNY, 496
variations in, 513–14
Gross fixed investment, 494
Gross investment, 497
Gross national income (GNY):
distribution of, 495–96
and equilibrium GNP, 510, 511, 512
and final demand, 495
gross business savings as percent of, 514
and measurement of GNP, 488, 489
multiplier effect, 526, 528, 530, 532–33, 534, 537
and net taxes, 517
and sectoral surpluses and deficits, 497–98
and value added, 491
Gross national product (GNP):
and balance of payments accounts, 744–45
constant vs. fixed dollar, 499, 501–2
and demand for money, 599, 602
and demand-pull inflation, 554–56
effect of interest rates on, 608, 611–20
effect on interest rates, 604–5, 606, 608
federal surplus or deficit as percent of, 475–77
and foreign trade sector, 518, 727–30
four-sector equilibrium, 516, 518–20
full-employment level, 619–20
long-term trends, 470–71, 498–99
measurement, 488–89, 490 *fig.*, 492, 494–95
and money supply, 597, 598, 614–16
and multiplier theory, 524, 525–28
and net taxes, 517
and planned demand, 530–34
and planned investment, 514–15
potential (*see* Potential GNP)
retrospective methods of calculation, 687
sectoral surpluses and deficits, 495–98
of Soviet Union, 811
and stabilization policy, 643, 644–45, 646, 647, 648, 649, 656, 657, 658
and stock market prices, 474
of Third World nations, 765, 766, 767–82
two-sector equilibrium, 510–11, 513, 516
and velocity of money, 602–3
Growth (*see* Economic growth; Population growth)
Growth strategies, of multinational corporations, 751 (*see also* Development strategies)

Haiti, 766
Harvard University, 231, 357
Hayek, Friedrich A. von, 43
Health maintenance organizations, 385
Health trends, 698
Heavy industry (*see* Capital intensity)
Heller, Walter, 626, 627 *fig.*, 667, 669
Henry George Society, 301
Hicks, John R., 42
Hidden unemployment, 629–30
High-employment budget, 646–49
Highway system, 707, 710–11
Hispanics, 406, 479
History, dialectic of, 803
Hoarding, 449
Holiday Inns, 344, 345 *fig.*
Holland (*see* Netherlands)
Homestead Act (1862), 693
Hong Kong, 772, 786
Horizontal mergers, 130, 135
antitrust policy, 257 *tab.*, 258, 264–66
as cause of firm dominance, 227
Household appliances, allocation of value study, 540–42
Households, defined, 19
Household sector:
in circular flow of goods and services, 23, 485, 486
conditions of efficient allocation, 358 *tab.*
consumption patterns, 104–5
and equilibrium GNP, 510–11
financial assets, 570–71
flow of funds, 573
functions in economic system, 18–19
income sources, 19, 34
maximizing behavior, 21–22
percent of GNY received by, 496
propensity to consume or save, 19–20, 34–35, 506–10
in Soviet economy, 812
stocks and flows, 51, 52 *tab.*
surpluses and deficits, 497, 498
see also entries under Consumer
Housing:
CPI changes 500–501
discrimination, 408–9
income effect on demand, 60
supply elasticity, 78
Human capital, 316–18, 691
Hume, David, 9

Imitation, phase of technological change, 352
Immature creditor, 750
Immature debtor, 749–50
Immigration:
and population growth, 688–92
of reserve army of unemployed, 799–800
Implicit costs, 25, 150
Imports:
competition from, 244
costs, 542–44
marginal demand propensity, 530
response to GNP changes, 524–25
tariffs and quotas, 720–26
see also Foreign sector; International trade
Import substitution, 786
Imputed costs, 150, 152

Incentives:
in China, 823, 824
for environmental protection, 433–35
and income support, 415–16
in Soviet Union, 816–17
taxation effects, 387–89
Income:
allocation of, and marginal utility, 113–16
of business sector, 513–16
and education, 316–19
effect of labor unions on, 326–27
of farmers, 460–62
household sources, 19, 34
as influence on demand, 58, 60, 61
international redistribution, 543–44
and investment demand, 34–35
by job type, 316 *tab.*
and output, 34
and preferences, 110–11
and price elasticity of demand, 67
requirements of enterprises, 20
and tax incidence, 411–13
see also Disposable income; Gross national income; Income distribution
Income distribution, 401–18
and allocative efficiency, 196
economic forces shaping, 405
effects of Reagan program, 682
as influence on demand, 58
and marginal productivity, 401
monopoly effects, 211–12
and public enterprises, 283
public finance effects, 389–92
and public policy, 409–17
in Soviet economy, 816–17
unfair, 367–68, 375 *tab.*, 402–9, 457
in U.S., 6, 50, 402–4
Income effect, 61, 308–9
Income elasticity of demand, 67–70
Income stabilization, vs. price supports, 460–61
Income statement, 138–39
Income support, and work incentives, 415–16
Income taxes (*see* Personal income tax)
Independent variable, 41–43, 46–47, 48
Indexing, 633–34
India, 393 *tab.*, 775
Individual demand, 103–22
and consumer surplus, 117–18
debate over microeconomic theory, 120–21
and derived demand, 120
marginal utility concepts, 107–9, 113–17
and market demand, 118–20
preferences and income effects, 110–11
rational choices, 105–7, 120–21
for scarce vs. free goods, 111–13
see also Demand
Induced changes, 524, 525–28, 530 (*see also* Multiplier)
Induced inventions, 353
Industrialization, forced, 809–10 (*see also* Development strategies; Economic growth)
Industrial mix, 226–27
Industrial Revolution, 250, 726
second, 733–34
Industrial sector:
aggregate concentration trends, 242
and antitrust policy, 259
largest firms, 128 *tab.*
supply of capital to, 338
in U.S. economy, 4–5
see also Business sector; Industries; Manufacturing sector
Industrial structure:
patterns and trends, 241–46
and public policy, field of, 222
in Third World, 772–74
Industrial unions, 322–23 (*see also* Labor unions)
Industries:
capital intensity effects, 77–78, 705
diversification of firms among, 126–27
with elastic demand and supply, 87
infant, 724, 786
labor-intensive, 704–5, 775–76, 786
life cycles, 275 *tab.*
minimum optimum scale, 171 *tab.*
news stories, 131
in 1972 input-output study, 492 *tab.*
primary, 4, 5 *fig.*
private, output in 1970s, 477–78
publicly owned, 283, 284 *fig.*
size of, and elasticity of supply, 77
Inefficiency:
allocative, 209–11, 212
in military weapons production, 438–40
X-level (*see* X-inefficiency)
Inelastic demand, 65–66, 67
Inelastic supply, 75
Inequality:
and discrimination, 406–9
in income distribution, 401–5
Infant industries:
argument for tariffs and quotas, 724
in Third World, 786
Inferior goods, 69–70
Inflation, 539–63
during Carter administration, 678–81
cost-push type, 540, 544, 551
due to defense spending, 658
defined, 539–40
demand-pull type, 540, 551, 552–57, 562
effect on banking industry, 577–78
and full employment, 619
due to large transfer payments, 659–60
macroeconomic concern, 36
monetarist view, 636–38
and money supply, 617–18
and multiplier process, 536–37
during Nixon administration, 672, 673–76
due to oil price rise, 543, 545–46
persistent (*see* Persistent inflation)
and potential GNP, 627, 629
price indexes, 499–502
and productivity, 549–51
and real interest rate, 656
role of labor unions, 544–46
statistical issues, 500–601

and unemployment, 11, 474, 546–49
Innovation:
to create and maintain monopoly power, 203
decision-making issues, 352–53
and education, 778
effects of monopoly on, 212
induced vs. autonomous, 353
phase of technological change, 352
process vs. product type, 352
return on, 342
Schumpeter's theory, 231
see also Technological progress
Input markets, 23, 289–304
economic rent concept, 299–302
equilibrium in, 298–99
supply and demand factors, 18–19, 290–98
and value of production, 302–3
Input-output accounting, 489–91
Input-output analysis, 357, 362
Inputs:
accounting vs. opportunity costs, 149–50, 151
categories, 20–21, 137
cost calculation, 153
cost of, and inflation, 540–49
efficient allocation, 361
key, and monopoly power, 204
least-cost combination, 171–73
prices of, and supply, 71, 72
relative value, 367
scarce, and elasticity of supply, 77
variable vs. fixed, 153
see also Input markets
In rem taxes, 386
Instability (*see* Economic instability)
Institutional investors, 130
Insurance industry:
largest companies, 129 *tab.*
public enterprises, 283
Insurance pooling, 385
Interbank deposits, 579–80
Intercept, 44, 45 *figs.*
Interdependence:
absence of, in monopolistic competition, 240
central concept of macroeconomics, 33, 34, 36
in oligopoly, 230–31
Interest effect, 611
Interest payments, 139, 140
foregone, 335–36, 337
Interest rates, 598–620
and asset values, 346–48
and business investment, 609–10
during Carter administration, 680
changes in, 606
and consumer demand, 608
and demand for money, 599–604
equilibrium rate, 604–6
and expenditures, 608–10
Federal Reserve actions, 606–8
and foreign portfolio investment, 752
high, effect on banking industry, 577
and internal rate of return, 334
Keynesian vs. monetarist views, 614–17
in the long run, 619–20
and monetary policy, 655–56
monetary feedback effect, 611–12
principle of equal advantage, 600–601
published listings, 131, 134 *tab.*
real vs. nominal rate, 617–18
and residential investment, 608–9
and return to capital, 338, 341–42
on time deposits, 568–69
usury laws, 95
and velocity of money, 603–4
Intergenerational choices, 454–55
Interlocking directorates, 254 *tab.*
Intermediate goods:
categories, 137
in circular flow, 486
cost of, and inflation, 540–42
imports, 487
input-output accounting, 489–91
price index (PPI), 499
Intermediate technology, 776–77
Internal funds, 337
Internal migration, 692–93, 810
Internal rate of return, 334–35
International Business Machines Corporation (IBM), 126, 171, 203, 224, 229
antitrust issues, 256, 257–58, 261–63, 269
International finance, 737–63
balance of payments accounts, 740–53
currency markets, 738–40
world payments system, 753–61
see also International trade
International Monetary Fund (IMF), 755–57, 761
International Telephone and Telegraph Company (ITT), 126, 244, 265 *tab.*, 266
International trade, 713–34
law of comparative advantage, 714–19
link to international finance, 738
Marxist view, 802–3
patterns of, 719–20
protectionism, 720–27
and U.S. economy, 727–34
of Third World nations, 785–86
see also Foreign sector
Interstate Commerce Commission (ICC), 251, 273 *tab.*, 274
Interstate Highway System, 707
Intrinsic value of money, 569
Invention:
induced vs. autonomous, 353
monopoly effects, 212
patent system, 203, 222, 353
phase of technological change, 352
see also Innovation
Inventories:
on balance sheet, 140
component of investment, 494
fluctuations in, 515
and unplanned investment, 511, 515
Investment:
cost of capital, 335–36
demand for, 34–35
and economic growth, 30
and federal budget, 665–67
and final demand, 494–95
foreign, Marxist view, 802–3
and interest rates, 609–10
internal rate of return criterion, 334
international, 730, 747–52
in job training and education, 316–19

marginal return (MRI) curve, 334–35
and personal saving, 34–35, 511–13
planned vs. unplanned, 510–11 (*see also* Planned investment)
present value criterion, 333–34 (*see also* Present value analysis)
profit-maximizing, 336–37
profits as a signal for, 142
and rate base of regulated utilities, 276, 277, 281–82
risk-return relationship, 338–41
as source of capital, 332–33
in Third World nations, 782–83, 786–88
Investment banks, 572–73
Investment goods, 486
Investor confidence, 473, 479
Investors, stock ownership, 127–30
Invisible Hand, 361, 364, 367, 370, 456
Iowa, 21, 457
Iran, 464, 474
Ireland, 688, 689, 691, 692
Irrigation, 701, 821–22
Italy, 284 *fig.*, 689, 782, 800

Japan, 4, 125, 792
and China, 818
current account, 745
environmental protection, 432–33
investment and growth rates, 30
investment in U.S., 751
life expectancy, 772
OPEC effects, 543
quota on car imports from, 724, 725–26
tax revenues/GNP, 393 *tab.*
trade effects of World War II, 729
trade patterns, 720
U.S. trade with, 718, 732 *tab.*, 733
Jevons, William Stanley, 8, 106
Jobs:
discriminatory policies, 406–8
pay rates, 315–16
selection of, 306–7
varieties, 315 *tab.*
see also Employment
Job training:
entry barriers, 324
return on investment, 316–19
Johnson administration, 11, 626, 667, 669, 671, 672, 678
Journal of Economic History, 12 *fn.*
Journal of Economic Theory, 12 *fn.*
Journal of Political Economy, 12 *fn.*

Kantrovich, Leonid V., 42
Kellogg Company, 227
Kennedy administration, 626, 627, 665, 667
Key inputs, 204
Keynes, John Maynard, 10, 11, 12, 476, 481, 524, 535, 681, 758
Keynesian economics:
and Bretton Woods Agreement, 758
historical attempts at stabilization, 665, 667–68, 671, 673, 678, 681, 682
ineffectualness in 1970s, 480
and liberal school, 13
vs. monetarist theory, 476, 598, 614–17, 653
as orthodoxy, 535
post-World War II influence, 11, 12
value of potential GNP gap to, 636
Keynesian Revolution, 667
Kinked demand curve, 236–39
Klein, Lawrence R., 43
Koopmans, Tjalling C., 42
Korea, 781 (*see also* South Korea)
Korean War:
as capitalist-socialist confrontation, 781
government spending, 474, 658, 683
inflation caused by, 472, 556
price controls, 95
Kung Bushmen, 737
Kuomintang, 818, 820
Kuznets, Simon, 42, 687

Labor:
component of factor services, 485
costs (*see* Labor costs)
Marxist view, 793–97
as production factor, 20, 137
productivity growth 349–51
relative value of, 302
unproductive, 800–802, 803
as variable input, 153
see also Labor market
Labor costs:
direct and indirect, 544
on income statement, 139
and least-cost production, 171–72
training costs, 316–19
union effects, 544–46
Labor force participation rate, 309–11, 677–78, 696–98
Labor-intensive industries:
defined, 148
in Third World, 775–76, 786
and U.S. economic growth, 704–5
Labor market, 305–29
artificial scarcity, 324
demand factors, 311–13, 324
equilibrium of supply and demand, 313–15
individual supply schedules, 308
job selection factors, 307
and marginal utility of work, 306–7
market supply schedules, 308–11
with monopsonist demand, 325–26
scarcity of talent, 319–20
union effects, 320–28 (*see also* Labor unions)
variations in pay rates, 315–16
Labor power, Marxist view, 410–11
Labor unions, 320–28
choice of goals and methods, 322–23
control over supply, 320–21
for craft and professional groups, 323–24
effect on demand, 334
effect on labor costs, 544–46
effect on workers' incomes, 326–27
leverage, 321–22
monopoly power, 204
and monopsonists, 325–26

response to persistent inflation, 558
Land, Edwin, 405
Land:
agricultural use, 700 (*see also* Farmland)
economic rent, 299, 301–2
production factor, 20, 21, 137
taxation of, 301–2
Land bank program, 461
Land reform:
in China, 819, 820–21
in Third World, 780
Law of comparative advantage, 714–19
Law of diminishing marginal returns, 27
vs. concept of economies and diseconomies of scale, 167
and educational investment, 318
and marginal revenue product, 291
and short-run productivity, 157–59
Law of diminishing marginal utility, 27, 107, 306–7 (*see also* Marginal utility)
Law of large numbers, 341
Lenin, V., 809, 820
Leontief, Wassily, 42, 357
Lewis, Arthur W., 43
Liabilities:
on balance sheet, 140
of commercial banks, 582–83
of Federal Reserve, 583–84, 592
of firms, 53
of households, 51
of major sectors, 52 *tab.*
Liberal economists, 13, 36
Life cycles:
human, and income distribution, 405
of industries, 275 *tab.*
Life expectancy, in Third World, 768–71 *tab.*, 772
Liggett and Myers Company, 255
Light industry (*See* Labor-intensive industries)
Linear equations, 41–44, 45 *figs.*, 46–47
Liquidity:
bank requirement, 580
of certificates of deposit and money market fund shares, 568–69
defined, 140
Literacy rates, in Third World, 768–71 *tab.*, 772, 779 *tab.*
Loan guarantee program, 395–96
Loans:
of banks, and creation of money, 584–91
of Federal Reserve, 584, 594
as financial asset, 570
interest rates, 342
as source of capital, 337
to Third World, 787
Local government:
spending trends, 394, 395
surpluses, 497
tax incidence, 411–13
tax revenues, 396 *tab.*
Location:
and choice of production technology, 148–49
of outputs, 138
see also Geography
Logarithmic scales, 48–49
Logic, 13, 47
London, 378, 694, 739
Long run:
defined, 153
equilibrium, 190–92, 197
productivity and costs, 163–73
supply, 193–94, 197
Loose oligopoly, 201 *tab.*
concentration ratio, 230
and tacit collusion, 235
Lorenze curve, 410
Los Angeles, 378, 428
Losses, minimizing, 186–87
Loss leaders, 216
Luck, 405
Luddism, 797

MacAvoy, Paul W., 464
Macroeconomics:
analysis of business cycle, 469–81
central concept, 33, 34, 506, 524
defined, 11, 470
major controversy, 614
principles, 33–37
Maine, 457, 709
Malthus, Thomas, 8, 448, 466 *fn.*, 688, 699
Management, and economies and diseconomies of scale, 168, 169
Managerial revolution, 127
Manchester School, 8, 302 *tab.*, 367
Manhattan Island, 21
Manufacturers Hanover Trust Company, 585
Manufacturing sector:
growth of, 695, 702–6
industrial structure, 242–43
large corporations, 126
1970s business cycle, 478
productivity gains, 350 *tab.*, 351
see also Business sector; Industrial sector
Maoism, 821, 822, 823, 824
Mao Tse-tung, 817, 818, 822, 823, 824
Marginal analysis, 14, 57, 106
Marginal benefits, 27 *tab.*, 136, 379–80
Marginal choices, and efficient allocation, 359
Marginal concepts, 27 *tab.*, 32–33
Marginal conditions, 25–26, 171–73, 359
Marginal cost:
and allocative efficiency, 195–96
calculation of, 159 *tab.*, 162–63
in competitive markets, 360–61
defined, 27 *tab.*, 182–83
and elasticity of demand, 294
equal to marginal revenue, 293, 297
equal to price, 185–86
of external funds, 337
of input vs. output, 290
under monopoly conditions, 206–9
and predatory pricing, 269
and profit maximization, 178–79
and regulated price levels, 272, 276, 277, 278–81
and supply curves, 183–90, 193–94
Marginal demand propensity (MDP), 530–34
Marginal effect (*see* Diminishing marginal effect)
Marginal product (MP):
calculation of, 154–58, 159 *tab.*
defined, 27 *tab.*

and least-cost combination of inputs, 171–73
and marginal cost, 162
and marginal revenue product, 290–92, 295
value of, 291, 296, 298
Marginal productivity:
and income distribution, 367–68, 401
and minimum wage law, 417
Marginal profit, 178–80
Marginal propensity, 27 *tab.*
Marginal propensity to consume (MPC), 508–10, 530, 534
Marginal propensity to save (MPS), 508–10
Marginal returns on investment (MRI), 334–35, 336–38 (*see also* Law of diminishing marginal returns)
Marginal revenue:
defined, 27 *tab.*
of dominant firm, 223
equal to marginal cost, 293, 297
with kinked demand curve, 238
and marginal revenue product, 290–92
under monopoly, 205–7, 209
in perfect competition, 181–82, 183 *fig.*, 295–96
and profit maximization, 178–80
and supply, 183–88
Marginal revenue product (MRP), 290–97
derivation, 290–92
and firm's demand schedule, 293–95, 311
in perfect vs. partial competition, 295–97
and profit-maximizing level of input use, 292–93
and value of individual inputs to production, 303
Marginal tax rates, 390
Marginal utility:
defined, 27 *tab.*, 108–9
diminishing, law of, 27, 107
of free vs. scarce goods, 111–13
microeconomic debate, 120–21
origin of concept, 106
preference and income effects, 110–11
and prices, equilibrium levels, 113–17
and total utility, 107–8
validity of concept, 121
and variety of goods, 110–11
of work, 306–7
Marijuana, government controls, 97–98
Marketable permits, 434, 435
Market demand:
and individual demand, 118–20
with perfect competition, 181–82
with pure monopoly, 204–6, 207–9
see also Demand
Market equilibrium, 78–80
government intervention, 94–98
in input markets, 298–99, 300
in labor market, 313–15
in oil market, 90–91
Market exchange, defined, 22–23
Market power, defined, 200 (*see also* Monopoly power)
Market price:
concerns of microeconomics, 55–57
and consumer surplus, 117–18
defined, 22, 23
and demand, 58, 59–67, 110
effect of expectations, 58
equilibrium point, 78–80
and marginal utilities, 113–17
of selected commodities and retail items, 56 *tab.*
of stocks, bonds, and commodities, published listings, 131, 132–33 *tab.*
and supply, 71, 72–74, 75–78
see also entries under Price
Markets:
antitrust criteria, 256–57, 260–61
basic functions, 57
causes of failure, 364–70, 374–75
conditions of efficient allocation, 358–61
defined, 22, 57
disequilibrium effects, 32
effect of conglomerates, 244–46
elasticity effects, 86–91
as focus of economic analysis, 23
input vs. output, 18–19 (*see also* Input markets)
interferences with, 91–98
interrelatedness of, and ripple effects, 356, 361–64
microeconomic analysis, 10–11
minimum optimum scale in, 170–71
naturally competitive, 226
pull toward equilibrium, 356–58
recent trends in concentration, 241–44
relevant, 260–61
types, 23, 200, 201 *tab.*
in Walrasian theory, 10, 357
see also Demand and supply
Market share:
antitrust criteria, 257–58, 259–60
concentration ratio, 229–30 (*see also* Concentration)
of dominant firm, 222–23, 224 *fig.*
and innovation, 203
mergers to increase, 203
and minimum optimum scale, 170–71
in monopolistic competition, 240
and monopoly power, 202
of oligopoly, 229–30, 239–40
Market value:
defined, 23
determinants, 106
and market prices, 55–57
Marshall, Alfred, 8–10, 25–26, 43, 106, 355
Marx, Karl, 8, 9–10, 12, 302 *tab.*, 476, 481, 535, 797, 803
Marxism, 791–804
class structure of capitalism, 487, 797–99
commodities, 793
critique of Soviet economy, 817
direct and indirect labor, 793–94
foreign trade and investment, 803
vs. Keynesian and monetarist theory, 92, 97
labor, labor power, and surplus value, 794–97
and radical school, 13
reserve army of the unemployed, 799–800
in Soviet Union, 808
unproductive labor, 800–802

Massachusetts, 697 *fig.*
Mature creditor, 750–51
Mature debtor, 750
Maximizing behavior, 21–22, 25
Meade, James E., 43
Means, Gardiner C., 127
Measurement:
of demand and supply, 98–101
diagrams, 40–50
interpretation of numbers, 50–53
see also Econometrics
Medicaid, 404, 413, 415, 659
Medicare, 286, 404, 659
Medicine, entry barriers, 319, 324
Medum of exchange:
gold as, 753–55
money as, 567
Mellon, Andrew, 405
Mellon family, 403, 405
Menger, Carl, 10, 106
Mercantilists, 8, 302 *tab.*
Mergers:
antitrust policy, 254, 255 *fig.*, 256, 257 *tab.*, 258, 264–67
conglomerate, 245
and firm dominance, 227
main kinds, 130–35
and monopoly power, 203
Mexico, 689, 774
Michigan, 552, 693, 724, 725
Microeconomics:
competitive assumptions, 180
debate over utility theory of demand, 120–21
defined, 10–11, 24
distinguished from macroeconomics, 33
focus on market prices, 55–57
new problems, 11
principles, 24–33
"secret password," 150
unifying concept, 355, 357
Middle Ages, 94, 387
Middle East, 444, 462, 463
Midwest, 690, 691, 692, 694, 709
Military aid, 787
Military draft, 442–44
Military spending, 437–45
avoiding production waste, 438–40
efficient levels, and arms race dynamics, 440–42
fluctuations, 517
government contracts nd monopoly power, 204
long-term trends, 475
pricing problems, 281, 439–40
under Reagan administration, 683
ripple effect, 365
and stabilization policy conflict, 657–59
volunteer army, economic basis, 442–45
Minimum optimum scale (MOS), 170–71
Minimum wage laws, 99, 416–17
Mining industry, productivity growth, 350 *tab.*
Minorities:
discrimination against, 406–9
equal opportunity programs, 413–16
unemployment, 479
Missallocation burden, 211
Mississippi, 693, 694
Mobil Oil Corporation, 217
Monetarist theory:
influence on Carter administration 679–80
vs. Keynesian theory, 476, 598, 614–17, 653
and Reagan program, 681
recommended stabilization strategy, 636–38, 646
as response to new problems of 1970s, 11, 480–81
Monetary base:
Federal Reserve control, 591–92, 593, 594, 655
and interest rates, 605
Monetary feedback, 611–12, 645
Keynesian vs. monetarist views, 614–17
Monetary policy:
under Carter administration, 679–80
conduct of, 652–56
coordination with fiscal policy, 656–60
discretionary, 638–39
under Johnson administration, 669–71
under Nixon administration, 673, 676
see also Stabilization policy
Monetary rule, 637–38
Money, 566–70
creation of, 584–95
as debt, 569–70
demand for, 599–604
effect of inflation on, 556
flow in economic system, 23, 24 *fig.*, 34
functions, 567
high-powered, 592
"hot," 752
neutrality of, 619, 620
vs. real capital, 332
stocks and flows, 51–53
velocity of, 602–4, 614, 616, 617
see also Money supply
Money market funds, 568–69
Money supply:
base (*see* Monetary base)
and Bretton Woods Agreement, 756–58
components, 568–69
effects of changes in, 653
and gold standard, 753–55
and inflation, 617–18
and interest rates, 599–620, 655–56
M1, 568–69, 570, 571, 578, 603
M2, 568–69, 578, 582, 603, 636, 637 *fig.*, 655
monetarist vs. Keynesian view, 598, 614–17
real vs. nominal, 617–18
role of depository institutions, 584–91
role of Federal Reserve, 578, 579, 581, 583–84, 591–94, 655–56
and stabilization strategy, 636–38
Monopolistic competition, 201 *tab.*, 240–41
concentration ratio, 230
distinctive concept, 222
instances, 225 *tab.*
Monopoly, 199–218
advantages of, for common property resources, 457
and allocative efficiency, 196
antitrust policy, 254 *tab.*
bilateral, 325–26
cases of, 216–18
characteristics, 200–203, 204 *tab.*
creation and maintenance of, 203–4
effects, 204–16, 218
and export restrictions, 726

and income distribution, 211–12, 405
labor unions as, 320, 325–26
laws against, 251
natural (*see* Natural monopoly)
and natural resource use, 455
and patent system, 353
public school system as, 423–25
pure (*see* Pure monopoly)
Soviet government as, 813
varieties, 200
see also Monopoly power
Monopoly franchises, 203–4
Monopoly power, 218
antitrust policies, 249–70 (*see also* Antitrust policies)
creation and maintenance of, 203–4
effect on demand for inputs, 295–97
effect on profits, 211, 343
gradations of, 200
indications of, 202–3
as limit on efficient allocation, 368–69
and market failure, 368–69, 375 *tab.*
regulation and public enterprises, 271–87 (*see also* Regulation)
see also Dominant firms; Monopolistic competition; Monopoly; Oligopoly
Monopsony, 320, 325–26
Moody's Industrial Manuals, 138
Moody's Investor's Service, 132, 341, 600
Morgan, J. P., 250
Mortgages, interest rates, 608–9
Moscow Olympics, 726
Multilateral aid, 788
Multilateral trade, 719–20 (*see also* International trade)
Multinational corporations:
foreign investment, 750–51
foreign trade effects, 730
response to strikes, 321
Multiple deposit creation, 585–88
Multiplier, 523–37
and federal budget, 643–45, 665–66
and fluctuations in planned investment, 514
process, 524, 528–35
role in macroeconomics, 524–25
size of, 531–33, 534–35
and stability, 533
theorem, 524, 525–28
uses, misuses, and limits, 535–37
Municipal bonds, 600
Mutual funds, 572
Mutual savings banks, 572
Myrdal, Gunnar, 42

National Broadcasting Company (NBC), 255
National Can Company, 255
National Highway Traffic Safety Administration (NHTSA), 435 *tab.*, 436
Native Americans, 406, 479
Natural competition, 170, 274–75
Natural gas, price controls, 95
Natural monopoly, 170
basic sources, 282
and limits to efficient allocation, 368
regulation of, 271, 272–73, 275, 282
Natural resources, 447–67
agricultural economics, 457–62
basic concepts, 448–57
depletion of, 447–48
energy, 462–65 (*see also* Fuels)
environmental protection, 428–35
future of, 465–66
as inputs to production, 137
limits to efficient allocation, 369–70
main types, 21, 448 *tab.*
and productive capacity, 5
scarcity, 1–2
supply curve, 298
Natural scarcity, 319
Necessities, 67, 272–73
Negative income tax, 416
Negative slope, 44, 45 *figs.*
Negotiable debts, 570
Neoclassical economics, 8–12
efficient allocation, 355
focus on marginal choices, 25
incompleteness, 12
and liberal school, 13
and limits of competition, 196
marginal utility analysis, 106
and value of production factors, 302 *tab.*
Net business saving, 496
Net exports, 497, 518
Net fixed investment, 494
Net foreign investment, 730
Net free reserves, 607
Netherlands, 284 *fig.*, 393 *tab.*, 425, 688
Net income, 137 (*see also* Profits)
Net national product (NNP), 495
Net plant and equipment, 140
Net taxes:
built-in stabilizers, 645–46
in calculation of sectoral surpluses and deficits, 497
and distribution of GNY, 496
and equilibrium GNP, 517
and stabilization policy, 643–44
Net worth, of households, 51
New Economic Policy, 809
New England, 461, 695
New Jersey, 688, 689
New Orleans, 695, 708
Newspapers:
dominance and scale economies, 228–29
relevant market, 260–61
Newsweek, 598
Newton, Isaac, 48, 505, 523–24
New York City, 95, 378, 384, 403, 404, 568, 600, 649, 694, 739
New York State, 273 *fig.*, 425, 688, 693, 707
New York Stock Exchange (NYSE), 343, 344, 345, *fig.*, 480, 572, 808
New York Times, 68–69, 500, 671
New Zealand, 774, 785
Nixon administration, 374, 672–76, 678
Nobel, Alfred, 687
Nobel Prize, 42–43, 357, 598, 687
Nonprofit enterprises, 20, 135, 385
Normal curve, 49, 50 *fig.*
Normal goods, 69–70
Normative economic analysis, 4
North America, 719, 720
Northeast, 6, 692, 693, 704
Norwood, Janet, 501
Nuclear fuels, cartel, 463
Nuclear power plants, 369–70

Obsolescence, technological vs. economic, 352–53

Occupational Safety and Health Administration (OSHA), 435 *tab.*, 436–37
Office of Management and Budget (OMB), 625, 626
Official settlements, 741–42, 757
Ohlin, Bertil, 43
Oil industry:
and alternative fuel sources, 454, 462–64
economics of exploration, 464–65
largest firms, 128 *tab.*
OPEC cartel (*see* Organization of Petroleum-Exporting Countries)
price fixing, 464
Stanard Oil monopoly (*see* Standard Oil Company)
U.S. trade, 718, 731
Oil prices
government controls, 95
increases in 1970s, 11, 543, 545–46, 673, 676, 681, 731
market equilibrium, 90–91
recent levels, 56 *tab.*
ripple effects, 363–64
Okun, Arthur, 626, 627 *fig.*
Oligopoly, 229–40
arms race as, 441
central tendency, 235–36
collusion, 232–35
competition within, 232
concentration and leading firms, 229–30
demand curves, 236–39
distinctive concept, 222
economies of scale, 239
instances, 225 *tab.*
interdependence of firms, 230–31
types, 201 *tab.*, 231–32
Open market operations, 591, 592–93, 594, 612–13, 653
Open shop, 320, 321
Opportunity cost, 25, 26
basic concept, 149–50, 151
and economic profits, 151–53
of holding money, 601–2, 604
of internal funds, 337
of loan guarantee program, 396
and marginal cost, 182
of military draft, 442–44
and production-possibility boundary, 29
of public expenditures, 379, 383
and question of producing or not, 178
and supply curve, 297
Optimum rate of use, of natural reources, 449–57
Optimum scale, 194
minimum, 170–71
Organization for Economic Cooperation, 766
Organization of Petroleum-Exporting Countries (OPEC)
as a cartel, 234, 543, 726–27
economies of member nations, 774, 784
effect on oil prices, 90, 464, 543, 545–46, 681, 731
export restrictions, 726–27
formation of, 543, 731
and internationl redistribution of wealth, 543–44, 752
trade patterns, 720
U.S. trade with, 732 *tab.*, 733
Output:
changes in, 35
of enterprises, 138
fluctuations in 1970s, 477–78
and income, 34
level of, and variability of costs, 159
long-term growth, 349–51
lost, due to GNP gap, 629
markets, 18–19
as monopolist's choice, 207
potential, 35–36
and pricing, 177–97
and productivity, 153, 549–51
profit-maximizing level, 178–80
setting, under perfect competition, 180–93
see also Gross national product

Parallel pricing, 234
Paramount Pictures case, 256
Partial competition, 201 *tab.*, 221–47
dominant firm, 222–29 (*see also* Dominant firms)
and marginal revenue schedule, 291, 295–96
monopolistic competition, 222, 230, 240–41
oligopoly, 229–40 (*see also* Oligopoly)
recent trends, 241–46
Patent system:
economic questions, 353
and firm dominance, 227
and monopoly power, 203
Pavarotti, Luciano, 319
Peak-load pricing, 278–81
Penn-Central Railroad, 245, 707
Pennsylvania, 688, 693, 747, 749
Pension funds, 581 *tab.*, 582
Perfect competition, 177–97
economic assumptions, 181
firm and market demand, 181–82
marginal revenue product schedule, 291, 295–97
short- and long-run equilibrium, 190–93
short-run supply curve, 183–88
Persistent inflation, 557–62
cost of, 632–35
costs of stopping, 630–32
defined, 540
monetary factors, 618, 652–53
price and wage controls, 635–36
Personal income tax, 386
and income distribution, 410, 411
"negative," 416
1964 reductions, 667
1968 surcharge, 671–72, 673
Reagan program, 682
Soviet equivalent, 816
trends, 396
and work incentives, 387–89
see also Tax incidence
Personal saving (SP):
and investment spending, 511–13
in 1970s, 634
propensity for, 508–10
Personal taxes, 386
Petroleum (*see* Fuels; Oil industry)
Philadelphia, 403, 694
Phillips, A. W., 546
Phillips curve, 546–49
with demand-pull inflation, 554
with persistent inflation, 557, 559, 560–62, 630–31
Photocopier market, 263–64
Physical capital, 331–33
Physical laws, and economies and diseconomies of scale, 168, 169
Physiocrats, 8, 302 *tab.*
Pittsburgh, 378, 428, 478, 695

Planned deficit, of business sector, 513
Planned demand:
effect of tax rates and unemployment insurance, 532–33
federal budget effects, 643–44
and multiplier theory, 525–28
response to change in GNP, 530–34
shifts in, effect on interest rates, 606
Planned investment (I):
and equilibrium GNP, 510–13
and marginal demand propensity, 530
and planned demand schedule, 526–27
response to changes in GNP, 524
variations in, 514–15
Planning-programming-budgeting (PPB) analysis, 380
Plantation agriculture, 780, 784, 785
Plato, 7
Pleasure (*see* Utility)
P. Lorillard Company, 255
Polaroid Corporation, 203, 227–29, 344, 345 *fig.*
Politburo, 813
Political influence, and natural resource use, 457
Pollution:
rules and fines, 384–85, 432
social cost, 378
social regulation, 428–35
see also Environmental protection
Population:
of China, 817–18
demographic changes, 695–99
and income distribution, 6
as influence on demand, 58
internal migration, 692–93
of Third World, 766, 768–71 *tab.*, 777–78
urbanization, 693–95
Population growth:
and future world resources, 465–66
Malthus's view, 8, 448, 688, 699
and production-possibility boundary, 29
trends, 688–92
Portfolio capital vs. real capital, 332
Portfolio investment, foreign, 748, 752
Positive economic analysis, 4
Positive slope, 44, 45 *figs.*
Post offices, public enterprises, 284 *fig.* (*see also* U.S. Postal Service)
Post-Revolutionary Society (Sweezy), 817
Potential GNP:
costs of exceeding, 630
costs of gap, 629–30
and high-employment budget, 646–49
and money supply, 636–39
Reagan view, 681
role in stabilization policy, 627–29, 638
Potential output, 35–36 (*see also* Potential GNP)
Poverty:
benchmark level, 6–7
in Third World, 765–66, 767–72
in U.S., 402–4
Precious metals:
market prices, 56 *tab.*
mercantilist view, 8, 302 *tab.*
published price listings, 131
Predatory pricing:
antitrust policy, 262, 269
by regulated utilities, 282
Preferences:
vs. actual market outcomes, 61
changes in, and marginal utility, 116
in household decision making, 19–20
and income, 110–11
as influence on demand, 58, 61, 104
of seller, during price ceilings, 96
for social goods, 377
variety of, 109–10
Prepaid goods, 113
Present value analysis:
and expected rate of return, 346, 347
of future uses of natural resources, 449–55, 456
of social goods, 382–83
use in investment decisions, 333–34
Preservation, vs. conservation, 449
Price ceilings, 95–96
Price controls:
arguments for and against, 635–36
interference with market process, 94–97
under Nixon, 674–76
Price cutting:
by oligopoly firms, 232–33, 234, 237–38
by regulated utilities, 282
Price discrimination
antitrust policy, 258, 268–69
by dominant firms, 224, 229
by electric companies, 218
due to monopoly, 213–16, 218
by regulated utilities, 272, 277, 279, 282
by Standard Oil, 217
by Xerox, 263
Price effect, on labor supply curve, 308–9
Price elasticity of demand, 63–67
and consumer surplus, 118, 119 *fig.*, 120
determinants, 67
estimating, precautions, 68–69
for selected goods and services, 100 *tab.*
Price fixing:
antitrust policies, 251, 254, 258, 267–68
of fuels, 463–64
by oligopoly, 232–33, 234–35
Price floors, 96–97, 99
Price indexes:
calculation of, 499–501
cost-of-living allowances based on, 633–34
GNP deflator as, 502
Price leadership, 234
Prices:
and allocative efficiency, 195–96
and average variable cost, 186–87
during Carter administration, 677, 678
changes in, effect on balance of payments account, 745–47
and demand, under perfect competition, 181–82
and determination of cost, 153
and economic choices of enterprises, 138

effects on, of dominant firms, 224
indexing, 633–34
as influence on supply, 71, 72
and least-cost production, 171–73
long-term trends, 498–99
and marginal cost, 185–86, 360–61
micro- vs. macroeconomic questions, 33
under monopoly, 207–8, 209
during Nixon administration, 672, 674–76
regulation of, 272–83
rigid, of oligopoly, 236–39
and scarcity of natural resources, 452, 455–56
and social values, 195, 196, 207
in Soviet economy, 815–16
see also Inflation; Market price; Pricing
Price signaling, 234
Price structure, regulatory issues, 272, 277
Price supports, 96–97, 460–61
Price takers, 180, 290
Pricing:
of military weapons, 281, 439–40
and output, under perfect competition, 177–97
predatory, 262, 269, 282
by public enterprises, 287
Primary industries, 4, 5 *fig.*
Primary products:
elasticity of demand, 780
importance to Third World, 785–86
Principle of equal advantage, 600–601
Principles of Economics (Marshall), 10, 26
Principles of Political Economy (Marshall), 106
Private enterprise:
aim, 20
corporations, 125–28
defined, 124
diversification, 126–27
mergers, 130–35
ownership and control, 127–30
small business, 124–25
see also Enterprises; Firms
Private good, vs. social good, 375–76
Private insurance funds, 581 *tab.*, 582
Private market system, 33, 455–57
Private sectors, cost of governance, 487–88
Process innovation, 352
Procter and Gamble Company, 140, 266
Producer price index (PPI), 499
Producers, share of sales tax, 92–94
Product (*see* Marginal product; Total product)
Product differentiation:
as entry barrier, 223
in monopolistic competition, 240
of oligopoly, 232
Product innovation, 352
Production:
basic input factors, 20–21
basic units, 20
capacity and growth factors, 5–6
capitalist mode, 793
diversity of techniques, 138
economic choices, 2, 136–37
efficient, 194–96
fluctuations in 1970s, 477–78
forces of, 798, 823
of households, 20
and income distribution, 6
of intermediate and investment goods, 486
least-cost method, 27, 171–73
relations of, 798, 820
supply and demand factors, 18–19, 71–72, 297–98
three basic questions, 2, 19, 24–25
valuing individual inputs to, 302–3
Production function, 154–55
Production-possibility boundary, 27–31
and comparative advantage, 715–17
macroeconomic equivalent, 35–36
in Soviet economy, 814–15
Production process:
choice of technology, 148–49
innovations, 352
inputs, 137–38
Productivity:
of agricultural sector, 457–58, 700–702
of capital, 332
of collective farms, 810, 820
and cost of output, 549–51
defined, 153
economies and diseconomies of scale, 167–71
in the long run, 163–73
Marxist view, 796–97
in the short run, 153–63
and technological change, 349–51
Product type, antitrust criteria, 260–61
Professional groups, 322, 323–24
Profitability:
correct measure of, 141
and monopoly power, 202–3
see also Profits
Profit maximization:
capital investment criteria, 334–37
as condition for equilibrium, 191
goal of private enterprises, 20, 22, 124, 127, 137
and input use, 290, 292–93
and least-cost technology, 148
by minimizing short-term losses, 186–88
under monopoly, 205, 207, 209
rules for, 178–80, 183, 184, 190–91
and short-run supply curve, 183–84
unnecessary in public enterprises, 283, 287
Profits:
accounting vs. economic, 138
after-tax, 139
and average variable costs, 186–88
calculation, 137
components, 342–43
distribution, 34
economic, 151–53
excess, 211, 224, 232–33, 272
joint, of colluding oligopoly, 232–33
and regulated price levels, 274, 277, 281
reinvestment, 337
as signal for investment, 142

see also Corporate profits; Profit maximization; Return on capital
Progressive taxes, 389–93, 410–13
Proletariat, 798, 813, 817
Property (*see* Assets)
Property income, 485
Property taxes, 301–2, 410, 413
Proportional tax, 410–13
Protectionism, 720–27
arguments for, 724–26
export restrictions, 726–27
tariffs and quotas, 721–24
vs. Third World export promotion, 786
Public choice, pinciple of, 32–33
Public enterprises, 20, 135, 283–87
Public expenditures:
alternatives to, 383–85
categories, 385–86
composition by level of government, 393–94, 395 *fig.*
cost-benefit analysis, 378–83
on education, 422–28
and income distribution, 413, 414 *fig.*
program types, 395–96
sources of funds, 383
tax loophole effects, 396–98
trends, 392–93
see also Government spending
Public finance, 373–98
economic concepts, 374–86
major patterns, 392–98
taxation issues, 386–92
see also Public expenditures
Public goods:
and consumer surplus, 118
costs, 366–67
marginal utility, 32, 113
Public policy:
economic concepts, 374–86
environmental, 428–35
and income distribution, 409–17
military spending, 437–45
worker and consumer safety, 435–37
see also Government policy
Public schools, cost-benefit issues, 422–28
Public sector, purposes, 374
Public utilities, regulation of, 249, 251 (*see also* Utilities)
Puerto Ricans, 479
Pure competition:
concentration ratio, 229
demand curve, 201
economic assumptions, 181
and input pricing, 290
vs. monopolistic competition, 241
vs. monopoly, 200, 201, 218
nature of, 180
vs. perfect competition, 181
Pure monopoly, 200, 201 *tab.*, 218
dominant firm likened to, 223
and market demand curve, 204–6
recent market share trends, 243–44
Pure public good, 286

Quantity controls, 97–98
Quantity demanded:
vs. demand, 61–63
equilibrium with quantity supplied, 78–80
by individuals, 117
and market price, 59–61
and price elasticity of demand, 63–67
Quantity supplied:
equilibrium with quantity demanded, 78–80
and market price, 72–73
vs. supply, 73–74
Quarterly Journal of Economics, 12 *fn.*
Queuing, 96
Quotas, in international trade, 720–27

Radical economists, 13
Radio Corporation of America (RCA), 244, 269
Railroad industry:
and agricultural growth, 705–6
early monopolies, 250–51
growth of, 709–10
life-cycle stage, 275 *tab.*
public enterprises, 284 *fig.*
Rate base, for regulated prices, 276, 277, 281–82
Rational expectations theory, 631, 683
Rationing, 96
Raw materials, as variable inputs, 153
Reagan administration:
antitrust policy, 256, 261, 263, 264,
economic program, 387, 392, 393, 416, 632, 681–83
safety regulations policy, 437
Reaganomics, 11
Real assets, 346
Real capital, 331–32
Real estate taxes, 387, 389, 391
Real GNP, 501–2
Recessions:
of 1958, 664, 668
of 1961, 664, 668
of 1970–71, 672–73, 759
of 1975, 676, 680, 731
of 1980, 677, 680
origin, 666
and unemployment, 6
Reciprocity, 737–38
Red Guards, 823–84
Regressive taxes, 389, 390–91, 410–13
Regulation, 272–83
applications, 272–73
background and evolution, 274–75
commissions, 273–74
cream skimming and competition, 282
decisions on price levels and structures, 275–77
and deregulation, 282–83
economic objective, 272
effects on cost, 281–82
marginal cost pricing, 276, 277, 278–81
problems and criticisms, 272
process, 275
Reagan program, 681–82
social, 428–37
Related goods, 72
Relations of production, 798, 820
Relevant market, 259–61
Rent:
controls, 95
CPI issue, 501
vs. economic rent, 299 *fn.*
Report of the Council of Economic Advisors, 627–28, 667, 675
Research and development (R&D):
by colleges and universities, 423

military, 438
and technological change, 351–53
Reserve deposits:
as commercial bank asset, 582–83
and control of money supply, 579
deficiency in, 588–89, 590
excess, 584–88, 589–91, 594, 606–7
as Federal Reserve liability, 584
and monetary base, 592
requirements (*see* Reserve requirements)
Reserve requirements:
changes in, 593–94
and interest rates, 606–8
and monetary policy, 653–54
and money supply, 584–91
prior to 1980 law, 580
under 1980 law, 581–82
reason for, 580
see also Reserve deposits
Residential investment, and interest rates, 608–9 (*see also* Housing)
Resource allocation:
capitalist vs. socialist, 782
and competition, 180
efficient, 355 (*see also* Efficient allocation)
monopoly effects, 209–11
in Soviet economy, 814–15, 816
Resources, wasted due to GNP gap, 629–30 (*see also* Natural resources)
Retail firms, largest, 129 *tab.*
Retail items, market prices, 56 *tab.*
Retained earnings:
and calculation of sectoral surpluses or deficits, 497
component of gross business saving, 513–14
effects, 342
on income statement, 139
and stockholder's equity, 140–41
Return on capital, 338–43
equalization of, 348
expected rates, 346–48
and interest rate, 341–42
and productivity, 332
profit components, 342–43
risk-return factors, 338–41
as success indicator, 141
see also Profits
Revenue Act (1964), 664, 671
Revenue and Expenditures Control Act (1968), 671
Revenue sharing, 394
Review of Economics and Statistics, 12 *fn.*
Ricardo, David, 8, 9, 23, 448, 688, 699, 714, 717, 719, 724
Ripple effects, 356, 361–64, 365
Risk:
and cost of capital, 337
defined, 338–40
diversification to reduce, 341
Riskless rate of return, 340, 342
Risk premium, 338, 340–41, 342
Rivalry, 181, 196, 202
Robber barons, 405
Robinson-Patman Act (1936), 254 *tab.*
Rockefeller, John D., 217, 229, 405
Rockefeller family, 403, 405
Rockefeller University, 405
Rome, 7, 8
Roosevelt, Theodore, 254
Rules:
to limit pollution, 432
and public policy, 383–85
Russia (*see* Soviet Union)
Rwanda, 772

Safety, social regulation, 435–37
Sales revenues:
in income statement, 138–39
of largest firms, 128–29
Sales taxes:
incidence, 387, 390–91
and income distribution, 410, 413
as in rem tax, 386
interference with market process, 91–94
Salvage operations, 284–86
Samuelson, Paul, 42, 598
Sandburg, Carl, 731
San Francisco, 479, 695, 709
Satisfaction (*see* Utility)
Saving function, 508, 509 *fig.*, 512
Savings:
and investment demand, 34–35
in Soviet economy, 812
in Third World, 776
see also Gross business saving; Personal saving
Savings and loan associations, 572
Savings deposits, 568, 581 (*see also* Bank deposits)
Savings institutions (*see* Thrift institutions)
Scale economies (*see* Economies of scale)
Scarce goods, 111–12
Scarcity:
and elasticity of supply, 77
of energy and fuel, 458–59, 462–65
as focus of economics, 1–2
of labor talent, 319–20
and least-cost production, 172–73
of natural resources, 448, 452, 455–56
and opportunity costs, 149, 150
and production-possibility boundary, 27–31
and production questions, 2–3
see also Demand and supply
Schools, public financing issues, 422–28
Schultz, Theodore W., 43
Schumpeter, Joseph A., 231, 260
Sears, 344, 345 *fig.*
Security markets, 572–73, 574 (*see also* Stock market)
Seller's preferences, with price ceiling, 96
Semifinished goods, 137
Services (*see* Goods and services)
Service sector, 4–5, 350 *tab.*
Shakespeare, William, 14
Shepherd, William G., 242 *fn.*
Sherman Antitrust Act (1890), 234, 251, 254, 255, 257 *tab.*
Shipbuilding industry, 284 *fig.*
Shopping Bag chain, 264
Short run:
costs, 158–63
defined, 153
equilibrium, 190–91, 197
vs. long-run efficiency, 166–67
productivity, 153–58
supply curve, 183–89, 197
Simon, Herbert A., 43
Simon, Julian, 466 *fn.*
Singapore, 772, 786
Single-parent families, 404
Size (*see* Economies of scale)

Skills:
differences in, 315–20
inadequate, due to discrimination, 406–8
and productive capacity, 5
as public benefit of education, 421
Slavery, 689, 696, 796
Slope:
of demand curve, 60, 64, 104, 110
of linear relationship, 44
of supply curve, 72–73, 75–77
Small business, 124–25
Smith, Adam, 8, 9, 12, 23, 168, 302 *tab.*, 361, 448, 559, 699, 713, 807–8
Social costs:
vs. external costs, 364–66
of public schools, 423
Social goods:
allocating spending among, 378–80
demand curve, 376–77
funding, 383
vs. private good, 375–76
privately funded, 385
and public policy, 374–76
Socialism, 8, 302 *tab.*, 781–82 (*see also* Marxism)
Social preference, 284
Social regulation, 428–37
Social Security system:
exempted from 1968 control act, 671
indexing of payments, 500
owner of national debt, 650
payments trends, 394
Reagan program, 682
and stabilization policy, 659–60
taxation trends, 396
Social services, 283
Social values, 195, 196 (*see also* Values)
South, 6, 404, 693, 694, 704, 709
South Africa, 737, 755
South America, 733, 766, 777, 786
South Asia, 733
Southeast, 6, 692, 704
South Korea, 733, 781, 786
Southwest, 403, 404, 689
Soviet bloc, 10
Soviet Union, 808–17
arms race, 441–42, 658
and China, 818, 819, 820, 822–23, 824
economic structure, 811–13
emergence, 803
history, 808–9
income distribution, 816–17
industrial development, 782, 809–11
Marxism, 792, 817
planning process, 813–15
prices, 815–16
private ownership, 781
Revolution of 1917, 692, 809
U.S. export restrictions, 726
wheat purchases, 553–54, 656
and Third World, 775, 787, 788
Spain, 284 *fig.*, 800
Special Drawing Right (SDR), 757, 761
Specialization:
colonial system, 780, 781
and comparative advantage, 715–19
and economies of scale, 168–69
transportation barriers, 717–18
Spillovers, 32
Sri Lanka, 772
Stabilization policy, 623–40, 641–81
built-in stabilizers, 517, 645–46, 666
coordination and conflict, 656–60
fiscal policy, 624–26, 642–49
goals, 626–36
historical record, 663–84
liberal vs. conservative view, 36
monetary feedback effect, 611–12
monetary policy, 612, 652–56
and multiplier theory, 525
and national debt, 649–52
and public sector, 374
strategies, 636–39
Stagflation, 474, 536–37
Stalin, Joseph, 809, 810, 811, 813, 817, 819, 820
Stalinism, 809–11, 817, 823
Standard & Poor's, 132, 341, 600
Standard Oil Company, 200, 217, 229, 254, 258, 262 *tab.*, 269
Stanford, Leland, 405
Stanford University, 405
State governments:
bank regulations, 578, 580, 581
spending trends, 394
surpluses, 497
tax incidence, 411–13
tax revenues, 396 *tab.*
Static analysis, 57, 149
Statistical methods (*see* Econometrics)
Steel industry, 284 *fig.*, 433
Steinbeck, John, 471
Stock:
accounting vs. market value, 140
asset values, 347–48
ownership and control, 127–30
see also Stock market; Stock prices
Stockman, David, 683
Stock market:
as a control system, 348–49
crash of 1929, 255, 470, 474
main parts, 344
relative impact, 566
see also Stock prices
Stock prices:
and asset values, 343–44, 354 *figs.*
fluctuations in 1970s, 479–80
long-term trends, 473–74
as performance index, 141–42, 348–49
published listings, 131, 132 *tab.*
Stocks, vs. flows, 51–53
Strachey, Lytton, 10
Strategy:
of multinational corporations, 751
of oligopoly firms, 230–31
see also Development strategies
Strikes, 321–22, 327–28
Structural unemployment, 552
Subsidies:
to public enterprises, 286–87
to public school system, 422–28
to transportation industry, 707, 710
Subsistence agriculture, 779
Substitutable goods, 58, 60, 294–95
cross-elasticity of demand, 70–71
and derived demand, 120
and marginal utility, 116
and market definition, 57
and price elasticity of demand, 67
Substitution effect, 60, 116, 312
Sumeria, 7
Sunk costs, 150, 151

Supply, 71–80
of capital, 337–38
diminishing marginal effect, 27
elastic, 86–87
elasticity of (*see* Elasticity of supply)
firms as basis for, 20
inelastic, 87–90
influences on, 71–72
in input markets, 297–303
interaction with demand, 78–80
long-run, 193–94
and the nature of costs, 147–74
vs. quantity supplied, 73–74
see also Demand and supply; Supply curve
Supply curve, 72–73
elasticity vs. slope, 75–77
graphing convention, 41–43
of individual laborers, 308–9
interaction with demand curve, 78–80
of labor market, 309–11
long-run, 193–94
under monopoly conditions, 206, 208–9
in perfect competition, 183–89
in pure competition, 180
shifts in, 189–90
upward slope, 72–73
see also Elasticity of supply
Supply side economics, 11, 681–83
Surplus, investment in financial assets, 570–71
Surpluses and deficits:
and equilibrium GNP, 512, 516
and financial markets, 566
in high-employment, 646–49
measurement of, 495–98
see also Deficits
Surplus value, 795–97, 799, 801–2
Survey of Current Business, 491
Sweden, 284 *fig.*, 393 *tab.*, 772
Sweezy, Paul, 817
Switzerland, 284 *fig.*, 393 *tab.*

T-accounts, 585
Tacit collusion, 234–35, 257 *tab.*, 268
Taft, William H., 255, 267
Taft-Hartley Act (1947), 320
Taiwan, 733, 786, 818
Talent:
and income distribution, 405
scarcity of, 319–20
and training, 318
see also Skills
Tariffs, 720–26
colonial system, 780
reduction in mid-1960s, 730
Taxes:
categories, 386
and distribution of GNY, 495–96
on economic rent, 301–2
effect of education on tax base, 421
on emitting pollution, 434–35
on imports (*see* Tariffs)
net, 517, 643–44
and products of government, 487
in Soviet economy, 816
see also Tax incidence; Tax rates; Tax revenues
Tax expenditures, 396–98, 423, 426
Tax friction, 389
Tax incidence, 386–87
and income distribution, 411–13
progressive vs. regressive, 389–93
and public expenditures on benefits, 413, 414 *fig.*, 415–16
Tax loopholes, 396–97
Tax rates:
incentive effects, 387–89
Johnson reductions, 667, 669
and planned demand schedule, 532–33
Reagan reductions, 682
surcharge of 1968, 671–72, 673
Tax revenues:
alternatives, 383–85
as percent of GNP, 393
trends, 396
for social goods, 383
Taylorism, 797
Technological obsolescence, 352
Technological progress:
and capital investment, 349–53
decision-making issues, 352–53
disadvantages, 796–97
and economic growth, 699–711
and natural resource use, 454
and patent system, 353
phases, 352
and production-possibility boundary, 29
and productivity, 5–6, 349–51
and pure competition, 196
sources, 353
see also Innovation; Technology
Technology:
agricultural, 458–59
and cost of production, 71–72, 148–49
and economic choices of enterprises, 138
and economies of scale, 227
least-cost, 164–67
and supply, 71–72, 189
Third World problems, 776–77
see also Technological progress
Teenagers, unemployment, 479
Telecommunications, public enterprises, 284 *fig.*
Telephone industry:
and economies of scale, 227
regulated price levels, 277, 279, 280 *fig.*
Tennessee Valley Authority, 284
Theory of Political Economy (Jevons), 106
Third-party effects, 321
Third World economies:
agricultural sector, 773–74, 779–80, 784–85
capital accumulation, 775–77, 782–83
capitalist vs. socialist development strategies, 781–82
dimensions of poverty, 767–72
education levels, 778–79
foreign investment in, 786–88
industrial structure, 772–74
population growth and control, 777–78, 783–84
relationship with First World, 780–81
significance of, 765–66
trade, 720 *tab.*, 732 *tab.*, 733, 785–86
uneven development, 774
Thrift institutions, 571 *tab.*, 572
reserve requirements, 581–82
Thurow, Lester, 260
Tight oligopoly, 201 *tab.*
antitrust policy, 258
concentration ratio, 230
price fixing cases, 267–68

recent market share trends, 243–44
strategy requirement, 230–31
tacit collusion, 235
Time, Inc., 344, 354 *fig.*
Time deposits:
in commercial banks, 571
money supply component, 568–69
reserve requirements, 581, 582
in thrift institutions, 572
see also Bank deposits
Time discounting (*see* Present value analysis)
Time series, 41, 48–49
Tinbergen, Jan, 42
Tobin, James, 43
Tolstoy, Leo, 767
Total cost (TC), 136–37, 159–60, 164
Total factor productivity (TFP), 349–51
Total fixed cost (TFC), 159 *tab.*, 160
Total product (TP), 154–58, 159 *tab.*
Total revenue, 63, 66, 136–37
Total utility, 107–8
Total variable costs (TVC), 159 *tab.*, 160
Toyota, 229
Trademarks, 203, 223, 27
Transaction deposits, 58[illegible]–82
Transfer earnings, 299–300
Transfer payments:
and distribution of GNY, 496
and equilibrium GNP, 517
to farmers, proposal for, 461–62
government expenditures, 386
vs. government purchases, 495
and net taxes, 517
and stabilization policy, 644, 659–60
to Third World, 787–88
trends, 394–95
Transformation curve, 715
Transportation cost:
barrier to trade, 717–18
and economies of scale, 227
Transportation industry:
development, 705–11
productivity growth, 350 *tab.*
Truman, Harry S, 624
Trusts, 250, 251
Turnover tax, 816
Unemployment:
and business cycle, 477–78
categories of, 552–54
during Carter administration, 676–77, 678
disguised, 775
effect on rate of change in money wages, 546–49 (*see also* Phillips curve)
fluctuations in 1970s, 478–79
frictional level, 536
hidden, 629–30
with increased government purchases, 645
with inflation, 474, 554–56, 557, 560–62, 630–31,
long-term trends, 471–72
macroeconomic concern, 36
Marxist view, 799–800
micro- vs. macroeconomic questions, 33
and minimum wage, 99, 417
and money supply, 619–20
and multiplier process, 535–37
natural rate, 562
during Nixon administration, 672, 673, 676
positive and normative statements, 4
post-1973 rise, 11
and potential GNP, 627–26
and Reagan program, 681
and recession, 6
time series diagram, 49
Unemployment compensation, 533
Unfair competition, 204, 254 *tab.*
Unfair distribution, 367–68, 375 *tab.*, 457
Union Pacific Railroad, 709
Union shop, 320
United Automobile Workers, 724, 725
United Fruit Company, 229
United Kingdom, 393 *tab.* (*see also* Great Britain)
United Nations, 788
United Shoe Machinery, 227
United States:
agricultural role, 553, 774
aid to Third World, 787, 788
antitrust policies, 250–69
consumption patterns, 104–5
distinctiveness of regulatory system, 272
distribution of income, 6–7, 50, 402–4
economic growth, 685–712
effect of Great Depression, 470–71
Gilded Age, 368
foreign trade, 720, 727–34
growth of federal government, 474–77
input-output analysis, 362–63
international money transactions, 740–61
largest firms, 128–29 *tab.*
merger boom, 130
number of private firms, 124
number of stock owners, 127
oil production, 462
productivity growth, 349–51
public enterprises, 283–86
role of corporations in, 124, 126–27
school enrollment, 422 *tab.*
tax incidence study, 411–13
topsoil losses, 459
trends in competition, 241–44
trends in public expenditures, 393–94
see also entries under Government; United States
United Technologies, 126
Unit labor costs, 549–50
University of Chicago, 476, 598
Unplanned investment (UI), 510–11
Urbanization:
in the U.S., 693–95
in the Third World, 772
U.S. Army, 383, 442–45
U.S. Bureau of Labor Statistics, 499, 500–501, 549
U.S. Census Bureau, 126, 229 *fn.*, 242
U.S. Congress:
antitrust actions, 253–54, 259
banking laws, 578, 581
budgetary actions, 374
environmental protection, 429, 432
and Federal Reserve, 584
immigration laws, 692
regulatory actions, 274
stabilization policy, 624, 625–26, 665, 667, 669, 671, 673, 674, 676, 678, 681, 682, 683